Kodikologie und Paläographie im digitalen Zeitalter 4

Codicology and Palaeography in the Digital Age 4

Schriften des
Instituts für Dokumentologie und Editorik

herausgegeben von:

Band 11

Schriften des Instituts für Dokumentologie und Editorik — Band 11

Kodikologie und Paläographie im digitalen Zeitalter 4

Codicology and Palaeography in the Digital Age 4

herausgegeben von | edited by

Hannah Busch, Franz Fischer, Patrick Sahle

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Philipp Hegel, Celia Krause

2017
BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 4. September 2017.

SPONSORED BY THE

Federal Ministry
of Education
and Research

Diese Publikation wurde im Rahmen des Projektes eCodicology (Förderkennzeichen 01UG1350A-C) mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) gefördert.
Publication realised within the project eCodicology (funding code 01UG1350A-C) with financial resources of the German Federal Ministry of Research and Education (BMBF).

2017

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7448-3877-1
Einbandgestaltung: Julia Sorouri, basierend auf Vorarbeiten von Johanna Puhl und Katharina Weber; Coverbild nach einer Vorlage von Swati Chandna.
Satz: LuaTEX und Bernhard Assmann

Inhaltsverzeichnis – Contents

Anhänge – Appendices

Vorwort

Der hier vorliegende vierte Band der Reihe *Kodikologie und Paläographie im Digitalen Zeitalter*, zugleich elfter Band der Schriftenreihe des Instituts für Dokumentologie und Editorik (IDE), versammelt zum einen Beiträge aus der Tagungsreihe *Maschinen und Manuskripte*, die in den Jahren 2014, 2015 und 2016 im Rahmen des Verbundprojektes eCodicology[1] in Trier, Karlsruhe und Darmstadt stattgefunden hat. Andere Beiträge wurden durch gezielte Anfragen durch die Herausgeber hinzugewonnen. Alle Beiträge wurden einer internen Begutachtung im erweiterten Herausgebergremium (IDE und weitere Fachleute) sowie einer anonymisierten Begutachtung durch externe Fachleute unterzogen. Dieser Band stellt das Ergebnis einer erfolgreichen Kooperation zwischen der eCodicology-Forschergruppe, dem IDE und einer äußerst aktiven Forschungsgemeinschaft im Schnittfeld traditioneller Geisteswissenschaften und Informatik dar. Tagungsreihe und Publikation wurden durch die dreijährige großzügige finanzielle Unterstützung durch das Bundesministerium für Bildung und Forschung (BMBF) ermöglicht, dem wir hiermit ebenso wie den Initiatoren des Projektes Andrea Rapp (Darmstadt), Claudine Moulin (Trier) und Rainer Stotzka (Karlsruhe) unseren aufrichtigen Dank aussprechen und auf das gesamte Projektteam bestehend aus Danah Tonne, Swati Chandna, Oliver Schmid und Vera Hildenbrandt ausweiten möchten. Des Weiteren gebührt unser Dank natürlich allen beitragenden Autorinnen und Autoren für ihre professionelle Zusammenarbeit auch unter bisweilen knappen Fristsetzungen. Gleiches gilt für die Fachgutacher und -gutachterinnen: für ihre konstruktive Kritik, die zu einer wesentlichen Qualitätssteigerung einzelner Artikel beigetragen hat, möchten wir uns vielmals bedanken. Herzlicher Dank gebührt schließlich unseren unentbehrlichen Helferinnen und Helfern: ganz besonderen Dank schulden wir Barbara Bollig (Trier) für zahllose formale und englischsprachliche Korrekturen und Julia Sorouri (Köln) für die Einbandgestaltung. Thomas Roesler (Köln) überprüfte alle URLs and archivierte am 22. Juni 2017 die referenzierten Webseiten soweit als möglich im Internet Archive (https://archive.org/). Bernhard Assmann (Köln) bewältigte erneut alle technischen Finessen der Drucklegung. Die redaktionelle Mitarbeit von Celia Krause (Darmstadt) und Philipp Hegel (Darmstadt) erstreckte sich auf alle wesentlichen Entwicklungsstufen dieses Bandes. Möge er einer interessierten Leserschaft zum Nutzen und zur Freude gereichen.

Köln und Trier im Juni 2017, die Herausgeber

---

[1] Dieses Forschungs- und Entwicklungsprojekt wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1350A-C gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

Preface

This book is the fourth volume in the series *Codicology and Palaeography in the Digital Age* as well as the eleventh volume of the publication series of the Institute for Documentology and Scholarly Editing (IDE). It represents the proceedings of the conference series *Machines and Manuscripts* organised within the frame of the collaborative research project eCodicology[1] held in Trier, Karlsruhe and Darmstadt from 2014-2016, whilst other contributors have been directly approached by the editors. All articles have undergone both an internal reviewing process by members of the IDE and editorial board and a blind peer reviewing process involving further experts from the field. This volume is the result of a successful collaboration between the researchers from the eCodicology project, members of the IDE and a very active research community working at the intersection of the fields of traditional humanities and computer science. The eCodicology project team and the IDE are grateful to the German Federal Ministry of Education and Research (BMBF) for enabling not only the conference series *Machines and Manuscripts* but also the realisation and publication of the present volume with their financial support. We would also like to thank the principal initiators of the project, Andrea Rapp (Darmstadt), Claudine Moulin (Trier), and Rainer Stotzka (Karlsruhe) as well as the whole project team consisting of Danah Tonne, Swati Chandna, Oliver Schmid and Vera Hildenbrandt. We want to thank all contributors for their professional co-operation which made the quick and smooth realisation of this publication possible. The same applies to all expert reviewers, we and the authors are thankful for the constructive feedback which helped to significantly raise the quality of the content of this volume. Our heartfelt thanks also go to indispensable supporters, especially to Barbara Bollig (Trier) for countless formal suggestions and language corrections as well as Julia Sorouri (Cologne) for designing the cover. Thomas Roesler (Cologne) verified all URLs and archived the referenced websites as far as possible in the Internet Archive (https://archive.org/) on 22 June 2017. Bernhard Assmann (Cologne) once again managed to cope with all technical intricacies of the print. Celia Krause (Darmstadt) and Philipp Hegel (Darmstadt) collaborated actively on all editorial decisions and steps regarding this publication. May it be the source of great avail and joy for its interested readers.

Cologne and Trier, June 2017, the editors

---

[1] This research and development project was funded by the German Federal Ministry of Education and Research (BMBF), funding code 01UG1350A-C, and managed by the Project Management Agency at the German Aerospace Center (PT-DLR).

Einleitung

Introduction

Kodikologie und Paläographie zwischen Geisteswissenschaften und Informatik

Andrea Rapp, Celia Krause, Philipp Hegel

Zusammenfassung

Digitale Kodikologie und Paläographie, wie sie sich in der Tagungsreihe *Maschinen und Manuskripte* des Projekts *eCodicology* präsentierten, werden üblicherweise in interdisziplinären Projekten und multidisziplinären Forschergruppen realisiert. Ein solches Design hat Konsequenzen für die digitale Kodikologie und Paläographie. Unsere Einführung beschreibt und systematisiert, wie mit dieser Situation bei den in diesem Band präsentierten Vorhaben umgegangen wird. Dabei zeigen sich Momente, die als ›zusammengesetzte‹ oder ›ergänzende‹ Interdisziplinarität verstanden werden können.

Abstract

Digital Codicology and Paleography as they are presented at the conferences of the *Machines and Manuscripts* series organized by the *eCodicology* project are normally realized in interdisciplinary projects and multidisciplinary teams. This research design has consequences for the structure of digital codicology and palaeography. Our introduction describes and categorizes the ways in which this situation is managed in the research projects represented in this volume. Aspects of composite and supplementary interdisciplinarity can be found.

1 Hintergrund

Der vorliegende Band fußt zum Teil auf Beiträgen aus der Tagungsreihe *Maschinen und Manuskripte*, die im Rahmen des Projekts *eCodicology* in den Jahren 2014 bis 2016 in Trier, Karlsruhe und Darmstadt veranstaltet worden ist. Für die erste internationale Tagung konnten Referentinnen und Referenten aus den Bereichen Paläographie, Kodikologie, Diplomatik, Informatik und Bibliothekswesen gewonnen werden. Die zweite Konferenz war technisch orientiert und bot Expertinnen und Experten der automatischen Mustererkennung oder der Informationsvisualisierung Raum für einen interdisziplinären Austausch. Die Abschlusskonferenz mit dem allgemein gehaltenen Titel *Forschung mit Schriftquellen im digitalen Zeitalter* hatte ihren Fokus auf

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. VII–XVI.

computergestützten Verfahren für die Analyse von handgeschriebenen Dokumenten und gedruckten Büchern. Analoge und digitale Methoden wurden explizit gegenübergestellt, um einen offenen Dialog zwischen traditionell und digital Forschenden zu ermöglichen. Buchkundliche wie auch technische Aspekte wurden gleichermaßen beleuchtet und ihre Relevanz innerhalb der Digital Humanities aufgezeigt.

Das Forschungsprojekt *eCodicology* hatte die Layoutanalyse von Handschriften des europäischen Mittelalters zum Gegenstand. Ziel war es, automatisch Maße auf Handschriftenseiten zu erkennen. Ausgangspunkt waren etwa 170.000 um Metadaten ergänzte Bilddigitalisate aus dem Skriptorium der Benediktinerabtei St. Matthias in Trier. Grundlegende Elemente der Gestaltung wie Seitenfläche oder Text- und Bildanteile wurden in ihrer Ausdehnung, Anzahl und Position auf jeder Seite durch den Einsatz von Bildverarbeitung und Algorithmen zur Merkmalsextraktion ermittelt, die dann statistisch ausgewertet wurden, um Muster und Veränderungen innerhalb des Bestandes von St. Matthias aufzeigen zu können.

2 Zwischen den Disziplinen

Digitale Kodikologie und Paläographie stehen, wie schon ihre Namen verdeutlichen, nicht mehr ganz auf dem festen Boden geisteswissenschaftlicher Tradition. Sie schweben zumindest scheinbar zwischen den Medien und den Disziplinen. So verstanden passen sie gut zu jenem nicht mehr ganz neuen Schlagwort der Interdisziplinarität, das seinerseits interdisziplinär behandelt wird. Wer aber schwebt, hat mit der Schwere zu kämpfen. Disziplinen sind selbst soziokulturelle Gebilde, die sich im Laufe der Zeiten herausgebildet, gefestigt und institutionalisiert haben. Auch wenn man anerkennt, dass bestimmte Gegenstände, Fragestellungen und selbst Methoden nicht nur von einer einzelnen Disziplin behandelt, beantwortet oder angewendet werden, heißt dies nicht, dass ein Austausch von Ergebnissen und Verhandlungen über Verfahren problemlos ist. Es gibt gute Gründe, warum sich Disziplinen ausdifferenziert haben und es für den einzelnen Wissenschaftler oder die einzelne Wissenschaftlerin schwierig ist, einen hinreichend hohen Grad an Spezialisierung in mehreren Disziplinen zu erreichen. Diese Schwierigkeit variiert mit den Disziplinen, aber zwischen traditionell ausgebildeten Geisteswissenschaftlern und Geisteswissenschaftlerinnen einerseits und Informatikern und Informatikerinnen andererseits ist die Differenz zumindest nicht zu unterschätzen.

Auch an den Gegenständen des Buches und der Schrift haben verschiedene Disziplinen Interesse. An der Tagungsreihe *Maschinen und Manuskripte* waren Philologien, historische Wissenschaften, Musikwissenschaft und Informatik beteiligt, in einigen Beiträgen wird ferner deutlich, dass sich auch Physik und Chemie mit diesen Gegenständen beschäftigen. Interdisziplinäre Tagungen, Projekte und Sonderforschungsbe-

reiche bringen seit einiger Zeit Vertreter und Vertreterinnen dieser unterschiedlichen Disziplinen zusammen, um diesen Gegenständen näherzukommen.[1] Die Tagungen des Projektes *eCodicology* und der daraus entstandene Band wollen einige dieser Perspektiven auf das Buch und die Schrift vorstellen. Dabei sollen die disziplinären Grenzen nicht verwischt, aber jene interdisziplinären Schnittmengen betont werden, die sich zwischen ihnen ergeben.

Das Projekt *eCodicology* war selbst insofern interdisziplinär angelegt, als Vertreter und Vertreterinnen aus Philologie, Geschichte, Archäologie und Informatik zusammengearbeitet haben, um Methoden der Bild- und Layoutanalyse mit kodikologischen Interessen, statistischen Auswertungen und Visualisierungstechniken zu kombinieren. Aus disziplinärer und interdisziplinärer Sicht sind zwei Punkte von allgemeinerem Interesse.

Bekannt und sprichwörtlich ist zum einen, dass Geisteswissenschaften und Informatik nicht immer die gleiche Sprache sprechen. Um sich eindeutiger verständigen zu können, wurde ein hierarchisch angelegtes Glossar kodikologischer Fachbegriffe in SKOS angelegt. Es steht nun auch anderen Forschenden zur Verfügung, um zum Beispiel automatisch vermessene Bildbereiche genauer zu beschreiben.

Das Projekt ist zum anderen in einem engeren Sinn interdisziplinär, wenn damit gemeint ist, dass Arbeitsschritte, die in einer Disziplin vorgenommen werden, Resultate aus anderen Arbeitsschritten voraussetzen, die in der anderen Disziplin vorgenommen werden. Die digitale Bild- und Layoutanalyse, die in das Aufgabengebiet der Informatiker fällt, schuf die Voraussetzungen für die statistische Auswertung der quantitativ arbeitenden Kodikologen. Die Voraussetzung war nicht zwingend, aber das Auffinden und Zählen von einzelnen graphischen Elementen durch den Kodikologen wäre deutlich weniger effizient gewesen. In diesem Fall handelte es sich also nicht um eine mögliche, noch enger gefasste Form von Interdisziplinarität, bei der Arbeitsschritte, die in einer Disziplin vorgenommen werden, Resultate aus anderen Disziplinen sogar notwendig voraussetzen. Aber die Verzahnung von Arbeitsschritten und Forschungsergebnissen spricht wohl dafür, hier nicht mehr von einer ›nur‹ multidisziplinären Sicht auf einen Gegenstand zu sprechen. Es lässt sich jedoch nicht von Transdisziplinarität sprechen, wenn damit gemeint ist, dass die einzelnen Wissenschaftler Arbeitsschritte übernehmen, die traditionellerweise von Vertretern einer anderen Disziplin übernommen werden.[2] Die Zuständigkeiten für die einzelnen Arbeitsschritte

---

[1] Zu denken ist an die Sonderforschungsbereiche *Materiale Textkulturen* in Heidelberg und *Manuskriptkulturen* in Hamburg. Zu denken ist auch an Projekte wie *Digital Resource and Database for Palaeography, Manuscript Studies and Diplomatic (DigiPal)*, die nicht von den Geisteswissenschaften allein umgesetzt werden.

[2] Jürgen Mittelstraß sieht in der Transdisziplinarität die »Interdisziplinarität im recht verstandenen Sinne«. Vgl. Mittelstraß 1987, 156: »Sie läßt die disziplinären Dinge nicht einfach, wie sie sind, sondern stellt, und sei es nur in bestimmten Problemlösungszusammenhängen, die ursprüngliche *Einheit der*

blieben vielmehr klar den einzelnen Disziplinen zugeordnet.[3] Die Verwendung von Resultaten aus dem Arbeitsschritt einer Disziplin in einem Arbeitsschritt der anderen verlangt aber ebenso eine Übersetzung wie die Formulierung von Anforderungen der einen Disziplin an die andere. Interdisziplinarität als Zwischenform zwischen Multi- und Transdisziplinarität kann diese Abhängigkeiten vielleicht ganz gut bezeichnen.

Die Verbindung mehrerer, disziplinär zugeordneter Arbeitsschritte und die Übersetzungsarbeit an den Schnittstellen zwischen diesen Schritten kann begrifflich präziser gefasst werden. Das Ineinander der disziplinär bestimmten Arbeitsschritte kann als Ausdruck einer ›zusammengesetzten‹ Interdisziplinarität mit kooperativen Momenten an den Schnittstellen gesehen werden. Zusammengesetzt heißt dabei, dass verschiedene Fähigkeiten zur Lösung eines Problems verwendet werden.[4] Kooperativ heißt, dass an den Nahtstellen Teamwork erforderlich ist.[5] Explizit wird das Ineinandergreifen von Disziplinen im Aufsatz von *Inga Behrendt*, *Jennifer Bain* und *Kate Helsen* anhand der Rolle des Musikwissenschaftlers bei der Arbeit mit einem digitalen Neumen-Wörterbuch erklärt. *Nanette Rißler-Pipka* nimmt in ihrem Aufsatz indirekt Bezug auf interdisziplinäres Arbeiten, indem sie darauf anspielt, dass sich der Blick der Geisteswissenschaft und der Blick der Informationstechnik auf digitale Bilder wesentlich voneinander unterscheiden.

Die interdisziplinäre Tätigkeit im Projekt *eCodicology* umfasste methodische wie theoretische Aspekte. Theoretische Interdisziplinarität findet sich bei der Erstellung eines gemeinsamen begrifflichen Rahmens in *eCodicology*, um die Ergebnisse der Bildanalyse in kodikologische Termini zu übersetzen.[6] Darüber hinaus wurde die

---

Wissenschaft […] wieder her.« In seinem Sinne wäre im Fall von *eCodicology* daher nicht von »rechter Interdisziplinariät«, sondern eher von Multidisziplinarität zu sprechen. Vgl. Mittelstraß 1998, 32: »Interdisziplinarität in Form von *Multidisziplinarität* läßt alles Fachliche oder Disziplinäre, wie es ist; man rückt nur auf Zeit, und ohne die eigenen fachlichen oder disziplinären Orientierungen irgendwie zur Disposition zu stellen, zusammen.« Der in dieser Einleitung verwendete Begriff der Interdisziplinarität soll anzeigen, dass ein Vorhaben als Ganzes mehr als nur multidisziplinär sein kann, obwohl alle grundlegenden Arbeitsschritte selbst disziplinär zugeordnet bleiben.

3 Dies lässt sich auch als eine ethische Maxime interdisziplinärer Arbeit verstehen, wie sie Ian Hacking in einer persönlichen Stellungnahme zum Thema ausgedrückt hat. Vgl. Hacking 2010, 196: »Worauf es meiner Meinung nach ankommt, ist, dass aufrichtige und gewissenhafte Denker und Engagierte gegenseitigen Respekt für ihre erworbenen Fähigkeiten und natürlichen Talente aufbringen.«

4 Zur dieser und der folgenden Begrifflichkeit vgl. Julie Thompson Kleins taxonomische Zusammenstellung. Vgl. Klein 2010, 18: »The label *Composite ID* names another familiar practice – applying complementary skills to address complex problems or to achieve a shared goal.« »ID« steht dabei für Interdisziplinarität. Vgl. auch in anderer Begrifflichkeit Klein 2010, 19: »In *Shared ID* […] different aspects of a complex problem are tackled by different groups. They possess complementary skills, communicate results, and monitor overall progress.«

5 Vgl. Klein 2010, 19: »*Cooperative ID* requires teamwork«.

6 Vgl. Klein 2010, 20: »The outcomes [der theoretischen Interdisziplinarität] include conceptual frameworks for analysis of particular problems, integration of propositions across disciplines, and new synthesis based on continuities between models and analogies.« Die beiden Bereiche arbeiteten bei

Methode der digitalen Bildanalyse von der Kodikologie ›entliehen‹, um die Gestaltung mittelalterlicher Handschriften zu erfassen.[7] In einem eigens erstellten Computerprogramm ist das geisteswissenschaftlich motivierte Training der Algorithmen zur Bilderkennung ebenso Voraussetzung für die Bildanalyse wie diese für die anschließende kodikologische Annotation und quantitative Auswertung. Wenn die Adaption von Methoden in eine dauerhafte Dependenz mündet, kann von ›ergänzender‹ oder ›supplementärer‹ methodischer Interdisziplinarität gesprochen werden.[8] Die Dauerhaftigkeit besteht in diesem Fall vor allem in dem digitalen Werkzeug, das fortan anderen, ähnlich ausgerichteten Vorhaben zur Verfügung steht. Der Beitrag von *Rißler-Pipka* in diesem Band stellt eine solche Nutzung der im Projekt *eCodicology* entwickelten Software in einem anderen Kontext vor. Ein weiteres Beispiel für die Nachnutzung der entwickelten Software ist ein Gastprojekt des Berliner Sonderforschungsbereichs *Episteme in Bewegung*, in dem automatisch erkannte Marginalien in der handschriftlichen Überlieferung des aristotelischen *De interpretatione* nach verschiedenen Kriterien klassifiziert und statistisch untersucht werden.

Auch die übrigen Beiträge dieses Bandes bewegen sich zwischen verschiedenen Disziplinen, sind aber oft ihrem Gegenstand, ihrer Frage- und Problemstellung, manchmal auch ihrer Methode nach einer traditionellen Disziplin mehr oder weniger klar zuzuordnen. Dennoch gibt es Schnittpunkte, in denen wie bei *eCodicology* Resultate einer Disziplin zum Material einer anderen werden oder Disziplinen auf andere Weise miteinander verzahnt agieren.

3 Gegenstände

Eine Möglichkeit, die Beiträge dieses Bandes zu ordnen, besteht darin, nach ihren zentralen Gegenständen zu fragen. Trotz aller Übergänge, die sich fast notwendig ergeben, lassen sich doch kleinere Gruppen spezifischen Inhalts identifizieren.

Sammlungen: Sammlungen können sowohl geschlossene historische Gebilde sein, die tatsächlich einmal bestanden haben, als auch ›künstliche‹ Zusammenstellungen von Texten nach verschiedenen Gesichtspunkten wie der Gattung oder dem Ort ihrer Herstellung. In beiden Fällen liegt der Schwerpunkt weniger auf dem einzelnen Objekt als auf dem Zusammenhang zwischen einer Vielzahl ähnlicher Objekte. Bei digitalen Sammlungen handelt es sich hauptsächlich um retrospektiv digitalisierte Quellenbe-

---

eCodicology aber dennoch auf der Grundlage ihrer eigenen Annahmen und Modelle.

[7] Vgl. Klein 2010, 19: »The typical activity [in methodologischer Interdisziplinarität] is borrowing a method or concept from another discipline in order to test a hypothesis, to answer a research question, or to help develop a theory«.

[8] Vgl. Klein 2010, 19: »If borrowing becomes more sophisticated and an enduring dependence develops, the relationship becomes *Supplementary*«. Beide Kriterien, Gewandtheit und Dauerhaftigkeit, sind relativ. Wann genau sie erfüllt sind, bedarf der Klärung.

stände kultureller Gedächtnisinstitutionen wie Bibliotheken, Archive oder Museen. Zu digitalen Sammlungen gehören aber auch Forschungsdaten, die im digitalen Medium generiert wurden. Digitale Sammlungen sind notwendige Voraussetzungen für die computergestützte Verarbeitung und Analyse. *Hartmut Beyer*, *Jörn Münkner*, *Katrin Schmidt* und *Timo Steyer* erschließen frühneuzeitliche Gelehrtenbibliotheken und stellen in ihrem Beitrag Möglichkeiten einer visualisierenden Auswertung vor, *Matthew Driscoll* untersucht die Überlieferung der isländischen »Geschichten der alten Männer aus den Nordländern«, der ›Vorzeitsagas‹.

Handschriftenbeschreibungen: Die Beschreibung von Handschriften ist ein weiteres Themenfeld, das in den versammelten Aufsätzen abgedeckt wird. Für die Erstellung von elektronischen Handschriftenkatalogen greift man oft auf die Angaben aus den gedruckten Katalogen zurück, die über einen langen Zeitraum erarbeitet worden sind. Durch eine umfassende Digitalisierung von Beständen ist es nun möglich, computerunterstützte Verfahren auf große Handschriftenbestände anzuwenden und neue Daten zu gewinnen. Handschriftenbeschreibungen können so mit zusätzlichen Informationen zu bekannten Beschreibungskategorien, aber auch mit gänzlich neuen Datenkategorien angereichert werden. *Alberto Campagnolo*, *Erin Connelly* und *Dot Porter* stellen das digitale Werkzeug VisColl vor, mit dem Lagen erfasst und dargestellt werden können. *Hannah Busch* und *Swati Chandna* beschreiben Werkzeuge, mit denen im Projekt *eCodicology* anhand von Digitalisaten mittelalterliche Handschriften automatisch vermessen und die Ergebnisse visualisiert werden können. Die Lagenbeschreibung und die Vermessung behandeln dabei zunächst jede Handschrift für sich, auch wenn im zweiten Fall herausgestellt wird, wie diese Daten anschließend für einen ganzen Bestand ausgewertet werden können. Im ersten Beitrag wird beschrieben, wie das Werkzeug den Kodikologen unmittelbar unterstützt und Außenstehenden anschließend einen Zugriff auf das Resultat erlaubt. Im zweiten Beitrag wird beschrieben, wie die Daten vom Computer gewonnen und mit bestehenden Metadaten kombiniert und verglichen werden. Auch wenn in diesen beiden Beispielen vorrangig Handschriften behandelt werden, sind die digitalen Techniken auch zur Beschreibung von Drucken geeignet.

Zeichensysteme: Ein drittes Themengebiet sind semiotische Systeme, die neben der Schrift das Buch ausmachen. *Rißler-Pipka* wendet das Werkzeug aus *eCodicology* an, um in lateinamerikanischen und spanischen Kulturzeitschriften der Moderne unterschiedliche Gestaltungprinzipien und -praktiken zu untersuchen. Dabei betrachtet sie insbesondere das Verhältnis von Bild und Text. *Behrendt*, *Bain* und *Helsen* versuchen die Aspekte der Bildverarbeitung und der Kodierung im Feld der Neumenforschung zu verbinden. Die digitale Behandlung mittelalterlicher wie moderner Handschriften und Drucke hat also nicht nur mit sprachlichem Text zu tun, sondern ebenso mit Illustrationen und Noten, die oft für die Beschreibung der Handschrift oder des Drucks relevant sind.

Schriftforschung: Definitionsgemäß legt insbesondere die Paläographie besonderes Augenmerk auf das Zeichensystem der Schrift. *Svenja Gülden*, *Celia Krause* und *Ursula Verhoeven* befassen sich in ihrem Beitrag mit der semantischen Modellierung und Visualisierung von Daten für eine digitale Paläographie des Hieratischen. *Torsten Schaßan* weist auf die Schwierigkeiten einer gemeinsamen Sprache für die Beschreibung von Schriften hin und skizziert Lösungsmöglichkeiten mit Blick auf Semantic Web-Technologien. *Bartosz Bogacz* und *Hubert Mara* beschreiben, wie sie mit XML-basierten Vektorgrafiken Keilschrift analysieren, *Enrique Vidal* stellt ein Programm zur Unterstützung der Transkription vor. Auch innerhalb der Analyse eines einzelnen Zeichensystems gibt es – ähnlich den beiden Projekten im Bereich der Handschriftenbeschreibung – unterschiedliche Grade, die Prozesse ganz oder teilweise einem digitalen Werkzeug zu überlassen. Der Grad der Unterstützung bei der Erkennung von Schrift und der Grad der Kontrolle werden je nach Gegenstand und Erkenntnisinteresse von verschiedenen Forschenden unterschiedlich bewertet und gehandhabt.

Datierung und Stemmatologie: Digitale Verfahren werden auch für die chronologische Einordnung der Forschungsobjekte eingesetzt. *Vincent Christlein*, *Martin Gropp* und *Andreas Maier* unterscheiden zwischen einem inhaltsbasierten und einem bildbasierten Zugang bei der Datierung von Handschriften und wenden eine automatische Methode an, indem sie Merkmale der Schrift auf Digitalisaten extrahieren und gruppieren. *Gábor Hosszú* wendet phylogenetische Methoden an, um die zeitliche Entwicklung ganzer Schriftsysteme zu untersuchen. Steht in dem einen Fall eher die Datierung des einzelnen Objektes im Vordergrund, so in dem anderen der Versuch, Ähnlichkeiten von Graphemen zu nutzen, um ihre Verwandtschaft zu beschreiben, auch wenn diese die Grenzen des einzelnen Schriftsystems überschreitet.

Materialität: Die Untersuchung der Materialität eines Buches am Digitalisat steht vor der besonderen Schwierigkeit, dass das digitale Objekt sich gerade materiell deutlich von dem eigentlich interessierenden Original unterscheidet. *Dariya Rafiyenko* beschreibt in ihrem Beitrag einen Weg, um digital mit der Materialität eines Palimpsests umzugehen und die *scriptio inferior* ohne großen technischen Aufwand sichtbar zu machen. *Campagnolo*, *Connelly* und *Porter* stellen in ihrem Aufsatz fest, dass bei der Präsentation von Digitalisaten oft Hinweise auf die physischen Eigenschaften von Büchern fehlen. Ihr Werkzeug VisColl soll bei der Beschreibung des Lagenschemas Abhilfe schaffen. Insbesondere Studien zum Layout von Schriftdokumenten können auch dazu beitragen, mehr über die Materialität dieser Quellen zu erfahren, wie in den Beiträgen von *Rißler-Pipka* sowie *Busch* und *Chandna* durchscheint.

4 Überschneidungen

Quer zu den oben genannten Gegenständen stehen einzelne digitale Verfahren, die in variierenden Kontexten unterschiedlichen Zwecken dienen. Die technischen Werkzeuge selbst sind durchaus vergleichbar, aber sie werden in unterschiedlichen Disziplinen zur Anwendung gebracht, um je eigenen Erkenntnisinteressen zu dienen.

Text- und Zeichenerkennung: Sowohl im *Optical Neume Recognition Project* von *Behrendt, Bain* und *Helson* als auch in der Anwendung von Handwritten Text Recognition (HTR) und Keyword Spotting (KWS) bei *Vidal* oder rootSIFT bei *Christlein*, *Gropp* und *Maier* werden Mechanismen zur automatischen Zeichenerkennung benutzt. Der phenetische Ansatz *Hosszús* setzt eine solche Zeichenerkennung voraus. Auch wenn diese Techniken in den vorliegenden Fällen Gegenstand der Informatik sind, zeigt die Einbindung von Trainingseinheiten, dass hier die geisteswissenschaftliche Beurteilung der Ergebnisse fest in den computerisierten Arbeitsablauf integriert ist.

Bildanalyse: Notwendige Vorstufe für die optische Erkennung von Schriftzeichen und Neumen ist die automatische Bildanalyse. Eine Bildanalyse kann sich auf 3D-Modelle oder Rastergraphiken von ganzen Schriftträgern, aber auch auf Vektorgraphiken von einzelnen Schriftzeichen oder Zeichengruppen erstrecken. Das Programm, das *Busch* und *Chandna* vorstellen und das von *Rißler-Pipka* auf ihr Zeitschriftenkorpus angewendet wird, konzentriert sich auf Elemente der Seitengestaltung. Auch bei der Bildanalyse dieser Art sind Geisteswissenschaftlerinnen und Geisteswissenschaftler in ein Training der Software involviert. *Rafiyenkos* ›Tracing‹ basiert zwar nicht auf einer vergleichbar automatischen Bildanalyse, aber auf einer definierten Routine zur Bearbeitung der Abbildung von Palimpsesten, um die verdeckte Schrift sichtbar werden zu lassen.

Kodierung: Neben der automatischen Gewinnung von Daten spielt die Kodierung als Modellierung von Information in mehreren Vorhaben eine Rolle. Die Kodierung nicht-alphabetischer handgeschriebener Schriftzeichen bildet im Aufsatz von *Gülden*, *Krause* und *Verhoeven* einen Schwerpunkt. Bei automatisch gewonnenen Daten wie im Projekt *eCodicology* sind die Speicherung der Ergebnisse und die Verknüpfung mit existierenden Beschreibungen Themen. *Driscoll* stellt in seinem Beitrag zu den isländischen Sagas einige Kodierungsbeispiele vor, darunter solche für Fragen der Seitengestaltung oder Textdichte. Im *Optical Neume Recognition Project* werden Kodierung und Zeichenerkennung aneinander ausgerichtet.

Normierung von Daten: Normdaten kommen in verschiedenen der dargestellten Projekten zum Einsatz. Zentral sind sie für das Anliegen von *Beyer*, *Münkner*, *Schmidt* und *Steyer*. Auch *Driscoll* nutzt normierte Daten für die Beschreibung und Auswertung der von ihm untersuchten Textgattung. Der Zusammenhang zwischen der Erforschung von Sammlungen und dem Anreiz, auf Normdaten zurückzugreifen, ist ebenso offensichtlich wie nachvollziehbar, da hier die Nutzung von etablierten Standards, die

Verknüpfung verschiedener Objekte, das Auffinden von Querverbindungen sowie die Prozessierbarkeit und Interpretierbarkeit der Daten zentrale Anliegen sind. *Schaßan* vergleicht bestimmte kontrollierte Vokabulare für die Schriftklassifikation und stellt diese zum Beispiel Taxonomien und Ontologien gegenüber.

Mikro- und Makroanalyse: Die Normierung der Daten ist oft eine Voraussetzung für statistische Auswertungen, wie dies in *Driscolls* Aufsatz deutlich wird. Dabei wird – wohl meist von geisteswissenschaftlicher Seite – betont und angestrebt, die Auswertung größerer Datenmengen in Form von Makroanalysen mit Mikroanalysen oder Einzelfallstudien zu verbinden, wie dies *Beyer*, *Münkner*, *Schmidt* und *Steyer* nach der Erschließung der privaten Büchersammlungen vorhaben. Mit dem im Beitrag von *Busch* und *Chandna* vorgestellten Werkzeug CodiVis ist eine Verbindung von Makroanalyse zum gesamten Handschriftenbestand und Mikroanalyse zu einzelnen Kodizes oder einzelnen Seiten möglich.

Visualisierung: Ein typisches und oft gewünschtes Ergebnis statistischer Auswertungen sind statische oder dynamische graphische Aufbereitungen der Ergebnisse in verschiedensten Formen. Auffällig ist dabei, dass dem Benutzer oft auch Möglichkeiten gegeben werden, mit den Visualisierungen zu interagieren oder diese mit Metadaten zu verknüpfen. So lassen sich sowohl bei CodiVis als auch bei VisColl Werte oder Gruppen von Werten auswählen, um so durch die Datenmengen zu navigieren.

Computergestützte Suche: Das sogenannte Information Retrieval dient der Suche nach komplexen Inhalten. *Vidal* präsentiert ein Modell, das zeigt, wie Indexierung und Suche ohne Transkription des Textes unmittelbar auf den Bildern selbst durchgeführt werden können. *Bogacz* und *Mara* stellen in ihrem Aufsatz eine graphische Suchmöglichkeit von Keilschriftzeichen vor, die auf Vektorgraphiken basiert und zugleich eine automatische Annotation gleicher Zeichen innerhalb des Textes ermöglicht.

Annotation: Eine weitere Form der Interaktion mit dem digitalisierten Material ist die Annotation. Der Ansatz von *Bogacz* und *Mara* beinhaltet ein entsprechendes Verfahren. Die Werkzeuge CodiVis und VisColl erlauben, automatisch oder selbst definierte Befunde mit einer Taxonomie zu verknüpfen. Auf diese Weise können entweder eine eigene Kategorisierung eingebracht und angewendet oder auch Daten mit extern definierten Standards wie Normdaten verbunden werden.

An verschiedenen Stellen wird deutlich, dass auch zwischen diesen Überschneidungen ihrerseits wieder Überschneidungen bestehen können. Ein Beispiel hierfür ist die Kombination von Kodierung und Zeichenerkennung. Einzelne technische Komponenten lassen sich miteinander oft auch in unterschiedlicher Reihenfolge zu einem Arbeitsablauf verbinden. Eine Verknüpfung mit Normdaten etwa kann über digitale Annotationen erfolgen und die Grundlage für statistische Analysen bilden. Diese Modularisierung digitaler Komponenten ermöglicht – bestenfalls – jeweils passende Kombinationen für verschiedene Arbeitsabläufe. Bei einigen digitalen Komponenten

ist zudem die Interaktion zwischen Geisteswissenschaftler und Software entscheidend. Nicht nur werden Programme oft so konzipiert, dass der Geisteswissenschaftler mit Darstellungen seiner Daten direkt arbeiten kann, in einigen Schritten wie den genannten ›Trainingseinheiten‹ sind informatische Kompetenz und handschriftenkundliches Wissen voneinander abhängig, um den Arbeitsschritt erfolgreich abschließen zu können. Die Software muss Ergebnisse liefern, die der Geisteswissenschaftler beurteilen und verwenden kann; die Ergebnisse der Software hängen aber auch von seiner Rückmeldung beim Training der Algorithmen ab. Supplementäre methodische Interdisziplinarität wie bei *eCodicology* ist also keine Seltenheit. Zusammengesetzte Interdisziplinarität mit kooperativen Momenten scheint sogar eher die Regel als die Ausnahme zu sein.

Bibliographie

CodiLab. <https://github.com/JochenGraf/CodiLab/blob/master/CodiKOS.html>.

DigiPal: *Digital Resource and Database for Palaeography, Manuscript Studies and Diplomatic.* <http://www.digipal.eu>.

Hacking, Ian. »Verteidigung der Disziplin.« In Jungert, Michael u.a. (Hrsg.). *Interdisziplinarität. Theorie, Praxis, Probleme.* Darmstadt: Wissenschaftliche Buchgesellschaft, 2010. 193–206.

Klein, Julie Thompson. »A taxonomy of interdisciplinarity.« In Frodeman, Robert (Hrsg.). *The Oxford Handbook of Interdisciplinarity.* Oxford: University Press, 2010. 15–30.

Mittelstraß, Jürgen. »Die Stunde der Interdisziplinarität.« In Kocka, Jürgen (Hrsg.). *Interdisziplinarität. Praxis—Herausforderung—Ideologie.* Frankfurt am Main: Suhrkamp, 1987. 152–158.

Mittelstraß, Jürgen. *Die Häuser des Wissens. Wissenschaftstheoretische Studien.* Frankfurt am Main: Suhrkamp, 1998.

Sonderforschungsbereich 933: *Materiale Textkulturen. Materialität und Präsenz des Geschriebenen in non-typographischen Gesellschaften.* <http://www.materiale-textkulturen.de>.

Sonderforschungsbereich 950: *Manuskriptkulturen in Asien, Afrika und Europa.* <https://www.manuscript-cultures.uni-hamburg.de>.

Sonderforschungsbereich 980: *Episteme in Bewegung. Wissenstransfer von der Alten Welt bis in die Frühe Neuzeit.* <http://www.sfb-episteme.de>.

Digitale Kodikologie

Digital Codicology

eCodicology: The Computer and the Mediaeval Library

Hannah Busch, Swati Chandna

Abstract

Through digitisation a large amount of mediaeval manuscript collection became publicly available, but the resources in time and human attention have not grown in proportion of digitised sources. Therefore, the question arises whether the computer can help to evaluate larger amounts of material like this. The project *eCodicology* has focused its research on the detection and measuring of the different layout features by using methods of pattern recognition for further analyses. The present paper gives insights into the developed software, SWATI – the Software Workflow for the Automatic Tagging of Images, and CodiVis, a visualisation framework for high-dimensional data sets, and how it can help the codicologist to explore the massive amount of heterogeneous datasets. The paper also focusses the various challenges, such as uncertain data due to irregularities and missing information in the manuscript's catalogues, as well as the accuracy of the image processing results.

Zusammenfassung

Durch die Digitalisierung sind zahlreiche Sammlungen mittelalterlicher Handschriften öffentlich zugänglich gemacht worden, jedoch sind weder die zeitlichen noch die personellen Möglichkeiten der Erforschung proportional dazu gewachsen. Daher stellt sich die Frage, inwiefern der Computer bei der Auswertung des Materials helfen kann. Das Projekt *eCodicology* hat seine Forschungsarbeit auf die Erkennung und Vermessung verschiedener makro- und mikrostruktureller Gestaltungsmerkmale der mittelalterlichen Seite gerichtet, indem es Methoden der Mustererkennung nutzt. Der vorliegende Artikel stellt die im Rahmen des Projektes entwickelte Software SWATI – Software Workflow for the Automatic Tagging of Images und CodiVis, ein Visualisierungsframework für hochdimensionales Datenmaterial, vor und erklärt, wie die entwickelte Software die Erforschung großer heterogener Datenbestände ermöglichen soll. Darüber hinaus richtet der Artikel sein Blickfeld auch auf die zahlreichen Herausforderungen die durch Unsicherheiten im Datenmaterial hervorgerufen werden sowie auf die Präzision der Ergebnisse der Bildverarbeitung.

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 3–23.

1 Page Layout and Mediaeval Manuscripts

A written page is more than text, it is not just a carrier of textual information, and the distribution of layout elements on the page can tell us more about the history of our written cultural heritage.

The page layout is defined as the collocation of rectangles containing graphical signs on the page surface of a book (Agati 2009, 219), the ratio between page and its content. The page layout aims to structure the codex and is designed according to the function of the text or book, to guarantee legibility. This is something everyone can notice by leafing through the codex. The appearance of a mediaeval book is very aesthetic, so it is hard to believe that it was realised by individual visual judgement, but research suggests the mediaeval artisans were artists rather than pure technicians. The question arises if they followed geometric rules, algorithms or a canon of proportions. This question has been the base of many layout studies concerning Latin and Greek manuscripts and it has been proven that at least in the most important scriptoria instructions had to be followed (see Maniaci 1995).

That the layout of the mediaeval manuscript page is not left to chance is proven by the existence of formulae of proportions as well (see Agati 2009 and Maniaci 1995). A formula of proportions can be defined as a coherent unit of standards, which – causing an organic bond between the different elements of the page – aims to extract the construction of a schema of ruling.[1] The formula must be un-ambiguous and universal, it must not give values but proportions between the different features of the page and it is sufficient to give essential parameters to obtain all layout features (Maniaci 1995, 17). The validation of a formula can only be proven if one applies a flexible approach with a tolerance range, not to forget that a manuscript is still an artisanal work.

Concerning the connection of geometrics and page layout, it is sufficient to observe the ratio between the two sides of the rectangle to understand if a notable rectangle is involved. Notable rectangles can be defined by proportions which converge the aesthetic ideals of antiquity and exhibit certain geometrical proportions between their long and the short side. Two of those antique visions of aesthetics are the Golden Ratio and the Pythagorean Theorem.

The theory is proven by certain recurring relations, like the relation between the height of the text block and the width of the page: h=L, or the width of the text block is equal to the page height divided in half l=H/2 (Agati 2009, 227ff.).

To verify such theories, analysis of large corpora of mediaeval manuscripts is required. Measuring hundreds and thousands of manuscript pages manually is a

[1] "Insieme coerente di norme che, istituendo un legame organico tra i diversi elementi della pagina, mira ad agevolare la costruzione di uno schema di rigatura univocamente definito" (see formulas of S. Remi Parisinus, lat. 11884, sec. IX., in Agati 2009, 219)

very time consuming undertaking and the error rate of human work increases with every page measured. The availability of digitised manuscripts offers the possibility to utilise computers to collect and process the data. The project *eCodicology*[2] is one attempt to analyse digital reproductions of mediaeval manuscripts with the help of computers by using methods of pattern recognition to take a closer look at the layout and perform statistical analysis of the newly gained data.

2 Introducing eCodicology

The idea of *eCodicology* was born during the digitisation project *Virtuelles Skriptorium St. Matthias* which digitised, reunited and published the manuscripts and fragments from the mediaeval library of the Benedictine Abbey of St. Matthias in Trier. Its basis is the idea of thinking further than just giving access to digitised manuscripts and catalogues.[3] For almost twenty years mediaeval manuscripts and other historical written documents have been digitised. Initially, digitisation focused on extremely important, famous, or rare manuscripts with the objective of making them accessible to the broad public and to ensure a better protection of the original. When high resolution scanners and digital single lens reflex cameras became more and more affordable, entire collections made their way into digital libraries.

New technologies and inventions have since been increasing the quality of the image data. It was time to take a next step and to rescue the digital collections from gathering dusty: digitised manuscripts can open new ways of research beyond better accessibility for researchers. The special research question of *eCodicology* focuses on generating new descriptive metadata by automatic analyses of digital images: is it possible to add missing or more precise information on the page layout in the catalogues by using the computer? And to which extent can these data help to support a historical research interest? To answer these questions, the project *eCodicology* tries to measure and analyse the page layout of mediaeval manuscripts by using the machine.

It has been the idea of *eCodicology* to establish a workflow for the automatic tagging of mediaeval manuscript layout features, including an algorithm library for pre-processing and feature extraction steps and transformation into the common format of the virtual scriptorium's database.[4] Furthermore, it experiments with the

---

[2] *eCodicology* is a joint research project of the Technical University of Darmstadt, the Trier Center for Digital Humanities and the Karlsruhe Institute of Technology. The project has been funded by the Federal Ministry of Education and Research (BMBF) under the agreement O1UG1350A-C from 2013-2016.

[3] The searchable database including the digital representations is available online via http://www.stmatthias.uni-trier.de and the TextGrid Repository (TextGrid).

[4] For the project *Virtuelles Skriptorium St. Matthias*, a MySQL database was set up. Since DFG-Viewer is used for the presentation of the facsimile's XML, information meeting the substandard METS was

exploration of these data by performing statistical analyses and by providing an interactive visualisation framework.

eCodicology follows the quantitative approach to codicology which was first developed by a group of French and Italian researchers in the 1970s. Instead of focusing their research on the description of single manuscripts, the group *Quanticod*[5] started to collect data for entire collections by building corpora and measuring similar features of the page layout. By manually collecting results of measurements and counting layout features on which statistical evaluations were performed, trends in manuscript production could be proven and displayed with graphic charts. Thus, it was possible to make statements about the character density on pages with a one or two column layout, about the significance of marginal space, and about temporary and regional tendencies concerning the *mise-en-page* of mediaeval manuscripts. Geometrical calculations could tell if the aspect ratio was influenced by norms like the Golden Ratio, well known from paintings, or the Pythagorean Theorem.

For the codicologist, the objective of working with the "masses" is to learn more about the materiality of manuscripts and their manufacturing process and to build a typology of manuscripts in a synchronic and diachronic perspective. For unknown reasons, the group of researchers stopped working on their projects just when computers developed more potential and, most importantly became affordable for research institutions and scholars.

3 SWATI – Software Workflow for the Automatic Tagging of Images[6]

In order to analyse a large quantity of digitised manuscripts one has to figure out how to prepare and to handle the image data, which, in the case of the St. Matthias scriptorium, are not less than five terabytes.

Therefore, the first goal of the project was to develop a complete workflow for automatic detection and tagging of layout features of mediaeval manuscript pages. Thus, high level interdisciplinary collaboration between humanist research and computer science was demanded. A first list consisting of properties which describe the page layout of a handwritten page containing all kinds of textual and pictorial elements such as highlighting, initials, decoration, changes of script had to be reduced

required. Descriptive metadata is stored according to the TEI P5 guidelines (TEI-C).

[5] Notably, Ezio Ornato, Carla Bozzolo, Denis Muzerelle, Dominique Coq. A collection of essays about their research has been published by Ezio Ornato in 1997.

[6] For a more detailed description of the *eCodicology* workflow, especially from the technical perspective, please see Chandna et al. 2015. The software will be available as a JAR file which can be executed directly from the command line. It will also contain plugins for ImageJ to test them with single images at a time. It will be published via the *eCodicology* project page

to initially three main features (page size, textual spaces, pictorial spaces) to ensure precision and quality of the automatically collected data (see fig. 1).

To extract the layout features of the manuscripts, different steps and applications of various image processing methods are necessary, starting with a pre-processing consisting of colour calibration, spatial calibration, noise removal, and scaling. Because image data themselves do not classify as codicological information, a schema was developed to translate them into such. However, the difficulties already begin when taking a closer look at the source language of the images.

First, one might have to deal with different resolutions due to the usage of different scanners or digitisation methods. Overhead scanners or constructions using DSLR cameras are most common. For the project *Virtuelles Skriptorium St. Matthias*, two different overhead scanners were used, one with a resolution of 300 dpi, the second with 400 dpi. This problem not only affects the resolution of the image, but also colour fastness. Different scanners have different colour spaces, which make digitised images dependent on scanner hardware. A software supposed to deal with any image must be able to adjust such variances. Using a colour checker and scale during the digitisation process is indispensable. Furthermore, images also have some noise that has to be removed or minimised by special filters. To analyse a larger amount of images, it is useful and possible to scale them down before processing them in order to reduce the time needed without distorting the results.

After these necessary pre-processing steps, an object segmentation can be performed. Object segmentation refers to the process which divides the image into its constituent objects and the background. The complexity of this process varies with factors like inconsistent intensity of the background, variations of intensity within the foreground objects, and clustering of foreground objects. The variations within the foreground might cause over-segmentation. For the *eCodicology* project, algorithms are trained to detect the borders of the page to measure the page size, textual space, and pictorial space. The training is done with the ImageJ Software (ImageJ) and MOA/WEKA (MOA), and utilises various machine learning algorithms, e.g. Bayes classifiers, Rules based classifiers, Tree classifiers, Lazy classifiers.

In the next step, the feature extraction, the relevant quantitative parameters are extracted using the images obtained in the segmentation step. The individual identification of segmented foreground objects can be done by labelling connected components. Pixels which belong to the same connected component are assigned the same label and, similarly, the pixels belonging to a different connected component are given different labels. Assuming the correct segmentation, the area of the manuscript page is extracted by counting the number of pixels corresponding to the foreground label. We extracted other features like page height and page width or text height and text width by reducing the foreground objects to arcs that are one pixel in width.

Figure 1: Example of a mediaeval manuscript page from Hs. 1108/55 4° (StB/StA Trier) showing the layout features page space, written space, pictorial space.

After pre-processing the images by calibration, filtering, and scaling, and successfully processing the results of the feature extraction, the data gathered can be included in the catalogue and used for statistical evaluation. This is the actual translation of image data into codicological information.[7] The results are saved in XML according to the TEI P5 (TEI-C) guidelines. For *eCodicology*, a TEI P5 conformant ODD-based metadata schema has been designed that allows storing metrical data in the manuscript description.[8] The resulting XML files are machine-readable and can be accessed by the software R (R-Project) or any other software to perform statistical calculations.

4 The Technique you use Influences the Result you get

The technical choice determines the possibilities of the scientific evaluation. This not only affects the quality of the scans. Some codicological features, such as watermarks, which have so far not been taken into consideration systematically in the catalogues for the library of St. Matthias, can only be detected on images produced with special techniques such as thermographic scans. Since no such scans are available for our corpus, we are not able to provide these extra information in the catalogue. But we can add more exact measurements for the page size of every single page, which is already a progress since traditional format data such as Folio, Quarto, and Octave can often be found in catalogues and the use of these descriptive data does not seem to be completely coherent. By using the computer, measurements of central tendencies and dispersion, maximum and minimum values in millimetre (or any other measuring unit) can be determined. It is correct, as Ezio Ornato stated in 1991, that quantitative codicology can be done with a simple sheet of graph paper (see Ornato 1991), but the larger the number of objects and variables, the more useful the application of a computer. The same is true for statistical evaluation. Furthermore, the computer is able to do this kind of work more precisely.

The structural data of a text correspond to the data extracted from digital images of mediaeval manuscript pages. In both cases, we see entities that are syntactically evaluated. A text consists of single elements such as chapters, words, and letters. Digital texts can even be enriched by paratexts or meta-information such as annotations or authority files. A digital image usually consists of single components or patterns that are machine readable based on their colour value, shape, or size. Humans see colour fields, shapes, symbols, or figures. Computers "see" pixels, hue, saturation, and brightness. Equal to text or language, the image of a manuscript page can be regarded as a sign system that can be processed and disassembled into its constituent components.

---

[7] For a more detailed description of the project's workflow see Chandna et al. 2015.

[8] ODD (One Document Does it All) files can easily be converted into various XML schema languages by using ROMA, a tool developed for generating customised TEI data.

But it is not only complex layout features which can be detected by a computer and used to answer codicological research questions. In the case of St. Matthias, fragments have been removed from the codices and collected in special archive boxes. However, in some cases it has not been recorded which fragments are related to which codex. Thus, two groups of fragments are given: one with recorded relations, and one without these data. Even apparently simple data such as height and width of page or text might help to reconstruct the original codicological connection.

Apart from first experiments with statistical evaluation mostly concerning the certainty of manual and automatic measurements, changes in page size, comparison of manuscripts written on paper and parchment, and the general book production of centuries , the project utilised a second approach to make the data talk: high-dimensional data visualisation.[9] With the help of CodiVis[10], a visualisation concept was developed to facilitate explorations of correlations in the abstract feature space of large sets of digitised mediaeval manuscripts (see fig. 2).

It combines two visualisation techniques in order to overcome the shortcomings of the single visualisation methods. In the first technique, manuscripts are clustered according to their bibliographic metadata and represented in a radial tree. This gives a quick overview of the whole data set. The polar node-link diagram was chosen over the Cartesian system because it combines the advantage of using space more efficiently while it has a pleasing aesthetic" (Heer et al. 2010, 64). In the second technique, bibliographic metadata are further linked to the macro- and micro-structural features in the parallel coordinate view, which is a relatively compact way to show many variables simultaneously. Interactive changes in the radial tree are automatically reflected in the parallel coordinate view. CodiVis consists of two major views: the manuscript explorer view, and the manuscript page explorer view. The former provides the users with an overview enabling access to the manuscripts and mean measurements of the layout features at a single glance. Users can select a subset of manuscripts which they want to explore and see the details in the manuscript page explorer view (see figs. 4-6). The latter allows the users to access the details of manuscript pages and measurements of layout features regarding each individual page. Both views help answering various domain specific questions such as "How is the distribution of manuscripts over the course of a particular century?" or "How did manuscripts develop over time with respect to the writing material?" The visualisation concept may show the potential of analyses by enabling quick exploration of "big humanities data".

---

[9] Human imagination reaches its limits when it comes to imagining more than three dimensions. To visualize high dimensional data sufficiently, visualisation or data have to be adjusted; therefore, special visualisation techniques have been designed.

[10] For a more detailed presentation of the CodiVis framework see Chandna et al. 2016 and 2017.

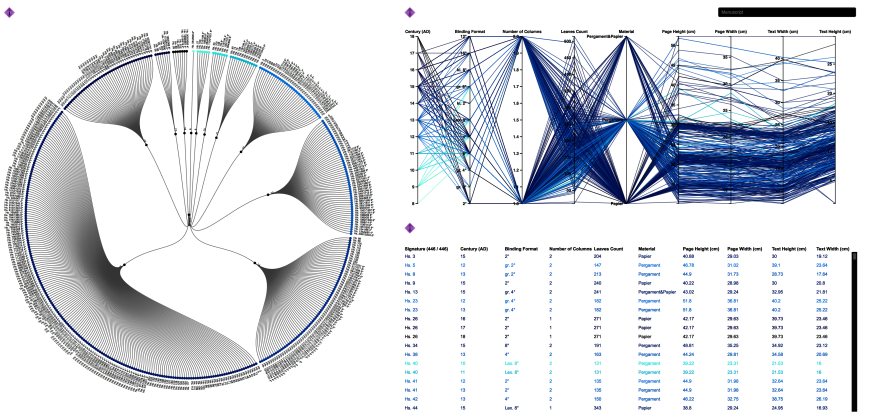


Figure 2: The image shows the main view of CodiVis with the list of all manuscripts in form of a radial tree on the left, the parallel coordinates with different parameters at the top right and at the bottom right the list of manuscripts with bibliographical metadata.

While statistical analyses can already reduce the necessary efforts for codicologists to extract knowledge, important information related to micro-level features might get lost as a result of research based on certain parameters and questions to the material.

The development of an effective workflow for image processing and interactive visualisation techniques has been one task of the project and is an important contribution to the future of codicological studies in the digital age since it allows the researcher to better handle digitised corpora.

However, realistically we also need to evaluate the accuracy of our results to see if they are worth all the effort. In the beginning of the project, we had an ambitious list[11] of codicological features hopefully to be detected automatically, which soon was reduced to areas of page size, textual space, and pictorial space (see fig. 1). The position of the textual space, which is recorded with coordinates (by combining TEI and SVG), can possibly give us information about the number of columns or glosses. In addition, the number of lines can be estimated. Irregularities such as highlighting, change of script, or initials influence the result. Currently, roughly 15 features – most of them assigned to one of the three main groups – are extracted on each page.

[11] The first list included different levels of text structure like headings, incipits/explicits, page title, page numbers, neumes, glosses, underlining, ruling, miniatures, initials, highlighting.

N.	Feature
1	Number of Pages
2	Mean Colour Value
3	Page Width
4	Page Height
5	Upper Left Corner Coordinates of Page
6	Relative Measurements of Page
7	Text Width
8	Text Height
9	Text Areas
10	Upper Left Corner Coordinates of Text
11	Relative Measurements of Text
12	Pictorial Width
13	Pictorial Height
14	Number of Pictorial Areas
15	Upper Left Corner Coordinates of Pictures
16	Relative Measurements of Pictures

Table 1: Layout features of the mediaeval manuscripts extracted by SWATI

	Bibliographic metadata	Values
1	Format	2°, 4°, 8°, 12°, 16°
2	Material	Paper, Parchment, Both, None
3	Century	8 AD, 9 AD, 10 AD, 11 AD, 12 AD, 13 AD, 14 AD, 15 AD, 16 AD, 17 AD, 18 AD

Table 2: Bibliographic metadata of the mediaeval manuscripts from St. Matthias database.

5 Challenges for Humanities Scholars: The Catalogue you have Influences the Result you get.

The stock of manuscripts of the library of St. Matthias is very heterogeneous, it is a collection bringing together the results of text production from the early 8th century to the 18th century, with different origins throughout Europe.

The lowest common property is that, at one point, they all became part of the St. Matthias library which only started to add a notice of possession to the codices in the second half of the 12th century.[12] With the occupation of Trier by French troops

[12] The first church in honor of the first bishops of Trier, St. Valerius and St. Eucharius, has its origins in

at the end of the 18th century, and the resulting secularization, most monasteries in the region were closed and the stock of the library of St. Matthias dispersed to various places around the world. Fortunately, the major part of the about 500 manuscripts remained in Trier and became part of the newly founded City library, today *Stadtbibliothek und Stadtarchiv Weberbach.* Today, a second, big part of the stock is part of the library of the episcopal seminary, the *Bischöfliches Priesterseminar Trier.*

In 2014, the project *Virtuelles Scriptorium St. Matthias,* funded by the DFG, completed the virtual reconstruction of the mediaeval stock of the monastic library; in this context, a digital catalogue database was set up and published and the roughly 440 codices which remained in Trier were digitised (Virtuelles Skriptorium). For the database the project team was able to revert to four catalogues describing the manuscripts of St. Matthias: in 1931, Josef Montebaur published a commented copy of the catalogue from 1530, Max Keuffer and Gottfried Kentenich catalogued the manuscripts stored at the city library – including the heritage of St. Matthias – between 1888 and 1931, the manuscripts of the episcopal library were registered by Jakob Marx in 1912. Unfortunately, none of the works contains detailed descriptions of each codex, neither regarding the content, nor the codicological description. The codicologist working with quantitative methods, thus, cannot refer to detailed preliminary studies. Furthermore, they have to deal with uncertain data. As mentioned above, the library stock of St. Matthias is very heterogeneous, including texts of various genres, proveniences, and centuries, often even bound together in one volume. In some cases, information about provenience, content, miniatures, and dating have been registered, but even though many of the manuscripts from the St. Matthias scriptorium or other scriptoria in the region have colophons, they have not been taken into account while cataloguing.

Almost every manuscript has been assigned a date of production; in most cases a century, in around 125 cases – especially for composite manuscripts – two or more centuries, and in a few cases even a more precise dating, the majority of those roughly 100 manuscripts is from the 15th century, the oldest manuscript is dated to the year AD 719. One might assume this to be a good basis to perform further analyses, but the dating is one of the project's major analytical uncertainties. It remains unclear what the date refers to: is it the point of production, acquisition, or registration in a catalogue. A parchment manuscript was probably not produced in the 18th century (although it is technically possible) and paper cannot appear before the 14th century. As seen in figure 3, the latest pure parchment codex is dated to the 16th century. In miscellaneous manuscripts with both materials, parchment can be found up to the

---

the second half of the fifth century. Between 970 and 980, the monastery became part of the Benedictine order, with this change at the latest a library must have been established.

18th century. Especially in the case of multiple-text manuscripts we do not have information dating single contents in the catalogue, sometimes there is no dating at all. A similar problem surfaces when taking a closer look at the assigned formats. In most cases in the St. Matthias database, information about the format was taken from the shelf marks, all codes of the codices from the city library end with an indication such as 4°, 8° or 16°. According to codicological description, this little detail should provide information about the number of times a piece of parchment has been folded and indicates the resultant number of leaves, but in this case most of the format indications have been added by the librarian. Therefore, it is more probable that the indication refers to the spine height of each book according to printing conventions to give information on which type of bookshelf in the depot a book can be found.[13] Thus, we cannot easily rely on that information either to perform analysis.

Working with the book or page size implies another problem: like in many other libraries, it was quite common to rebind the manuscripts and to cut them to a new standardized size in order to match the style of the institution. Cutting processes mostly have not been registered but taking a closer look at the book the cutting edges are clearly visible which makes it very difficult to perform satisfying studies of (original) book formats. The binding also swallows an indefinable part of the inner margin and can therefore be considered a major enemy of page measurements. Other than modern books, manuscripts do not have a title, instead, the incipit or first words were noted to the catalogue entry. Another information gap is the description of initials and miniatures. Not all of the texts are illuminated or contain miniatures, but initials were a common feature to structure the texts, starting from little highlighted letters to page-filling initial letters. Only a very small percentage of pictorial elements have been catalogued, detailed descriptions, measurements, and information about the exact location within the codex are missing. Therefore, we have no information about the text-to-image ratio in mediaeval manuscripts. To find and count images, one would have to leaf through all the books which, of course, got easier with every digitisation but is still a very time-consuming procedure. Applying SWATI on a digitised set of manuscripts will provide you with all these information by adding them to the metadata.

Thus, on the one hand, we can be lucky to have a catalogued corpus to work with, but, on the other hand, our catalogues only contain basic and often uncertain data. Hence, one goal of the *eCodicology* project was to discuss those problems and to develop a workflow to handle uncertainties. With the automatic approach

[13] The results of a comparison between format, translated into measurements in centimetre according to the guidelines of the *Deutsche Bibliothek*, today *German National Library*, and the actual format of the binding are presented in the final report of the project (eCodicology). The comparison is of course not exact because of missing references to mediaeval times where the size of a page was determined by type and size of the animal (skin) and the number of times the parchment had been folded.

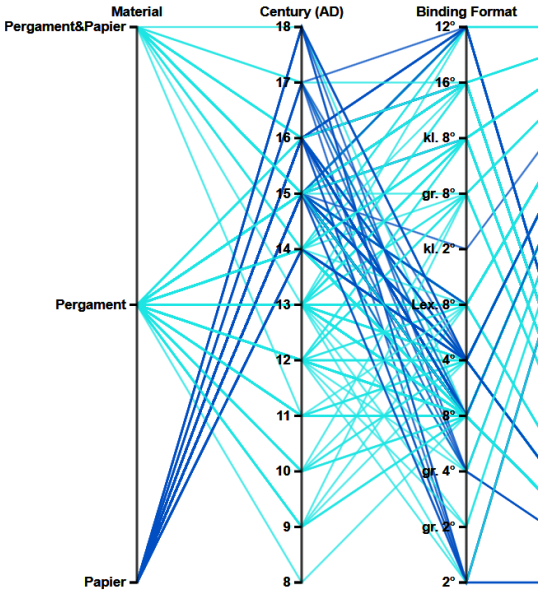


Figure 3: The image shows a detail from the parallel coordinate plot in CodiVis: the emergence of different materials and their dating.

from the computer science part of the project it was even more important to discuss uncertainties and to determine solutions. Format and automatically measured page size are two entirely different entities and will not be mixed but registered separately in the TEI P5 XML files without any loss of information.[14] For the storing of the manuscripts, the dating held further problems since a solution needed to be found on how to deal with manuscripts dated to two or more centuries. Even though structural data which separate the different contents in one manuscript have been generated during the digitisation process, neither human nor machine can easily define a date of production for each content. The question remains how to handle these uncertain data. As a solution, manuscripts with more than one dating in the TEI file are represented with multiple lines in the parallel coordinate plot.

Visualisation can help to find outliers as it shows certain patterns, for example for the page size and the main text areas. Significantly divergent formats in a time period can hint at the certainty of the temporal placement in the bibliographical

[14] A detailed description of the developed schema is described in the final report of the *eCodicology* project.

metadata. The same applies for questions about the writing material. Including information about a manuscript's content is rather difficult: as mentioned before, the more information the catalogue offers, the more questions can be answered by a quantitative approach. While digitisation enabled access to large collections of manuscripts and initiatives such as the *Text Encoding Initiative* (TEI-C) have developed standards to store bibliographic metadata in an interoperable format, most of the recent manuscript cataloguing projects have not kept in mind to provide their data in machine readable files. The non-existence of a multilingual controlled vocabulary for manuscript description to refer to is a serious problem.[15] The idea to face this problem within the project *eCodicology* was born due to the difficulty of interdisciplinary and intercultural communication in the project team, which consisted of scholars with different nationalities and from disciplines, namely computer science and various disciplines from humanities research fields. A bilingual list of terms was collected to simplify communication between the project members and transferred to a browser-based SKOS editor to build up a rdf-based codicological ontology (CodiKos).[16]

The quantitative codicologist cannot gain new insights without the support of the expert applying traditional methods. A closer collaboration is highly demanded in the future. Without being able to work with library contents annotated to higher standards, the building of larger corpora with the same parameters is impossible and limits the option of quantitative research.

6 Discussion of the Results: What did eCodicology Teach us so far?

At the preliminary endpoint of the project work[17], all 170,000 pages (440 codices) have been processed once, roughly 15 different features (see table 1) per page – 10.000 features per manuscript – were captured and saved in the metadata files, which makes a total of 2.5 million new entries. The processing of one codex page takes 3-5 minutes according to the complexity of each page. The newly gained metadata give – apart from information on e.g. the colour values – information about the page size and dimensions and place of the different text and image areas on the page. After the first run of SWATI we got satisfying results[18] regarding the page measurements

[15] For further information on the general difficulties of manuscript terminology across different languages see Jakobi-Mirwald 2009.

[16] The SKOS ontology was developed within CodiLab in collaboration with the project *SemToNotes*, which aimed at designing an image annotation tool for manual correction and semantical enrichment of automatic image annotation.

[17] The funding of *eCodicology* ended in April 2016. The software development is currently continued in the context of a Phd thesis at the Karlsruhe Institute of Technology.

[18] We decided on a 2,5% acceptance range for deviation, everything within that 2,5% is satisfying.

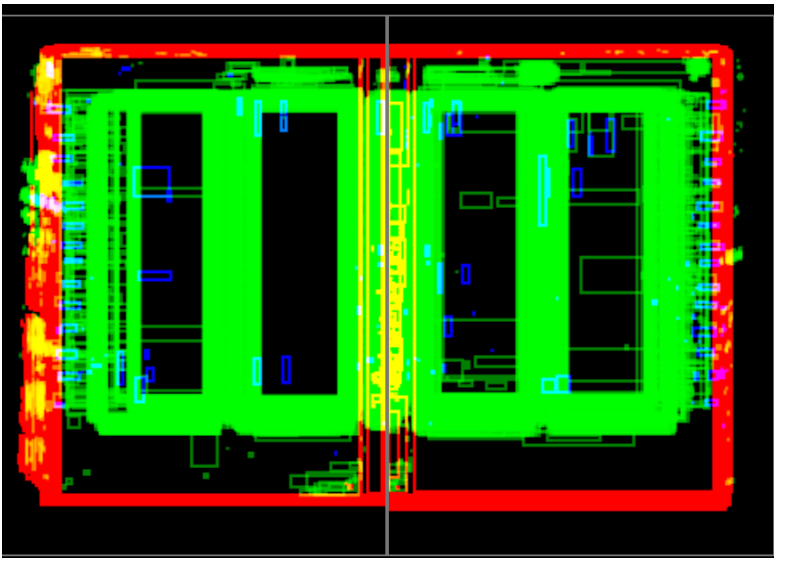


Figure 4: This figure shows the superimposition plot of page space measurements, textual space measurements, and pictorial space measurements

and main text areas. After processing the first couple of complete codices, manual measurements of the same sample were taken and compared to the automatic results. An acceptable range of 2,5% for the mean deviation was decided and both groups were within that range concerning the page height.[19]

Currently (January 2017), calculations and tests to define the segmentation quality are being performed. The feature extraction of image areas can still be considered to be in its infancy: extracting image areas is a very complex process as image areas may contain many areas too similar to the background of the page to be determined precisely and it needs more training before one can work with the results. To store the results of feature extraction correctly they need to be filtered since the machine defines irregularities as text or image space. After sampling and defining a minimum value for those areas and further training the algorithms, the results can be improved significantly. Visualisation of the annotated areas can support these processes (see fig. 4).

Looking at the data visualisation on page level, the detected zones can be reflected in the original image: red frames surround the page area, green frames the text areas, and blue frames the image spaces. The example (fig. 4 and 5) shows manuscript Hs. 68

[19] The case study has been presented at the annual conference of the DHd in Graz in February 2015.

Figure 5: This figure shows a manuscript montage plot. In this view, all the manuscript pages are combined into the composite view and the relative measurements extracted from the SWATI workflow are plotted on top of each respective page.

from the *Bibliothek des Bischöflichen Priesterseminars Trier*, a manuscript on parchment from the 12th century with a two columned layout and without illuminations. The two main text areas can be distinguished very well from misdiagnosed smaller areas which most probably are irregularities on the parchment. Little text areas in the upper right corner are most probably the page numbers added subsequently and the few oblong text areas on the outer margin are glosses. Large red areas left-aligned within the text fields are initials. The remaining detected zones can be marked as errors for further training. Training can improve the processing algorithms significantly. Statistical analyses can help to define a value for minimum areas of text and image to be ignored and location of an area on the page can also help to evaluate the correctness of a detected area: features very close to the binding are mostly neither text nor image. The accuracy of the measurements is also checked with confidence intervals of currently 20 pages to get a mean of the whole manuscript, if those twenty pages show accurate results, training and feature extraction is considered successful.

2.5 million features on 170,000 image scans is too big a number to handle manually for further analyses. Therefore, the visualisation framework CodiVis was developed

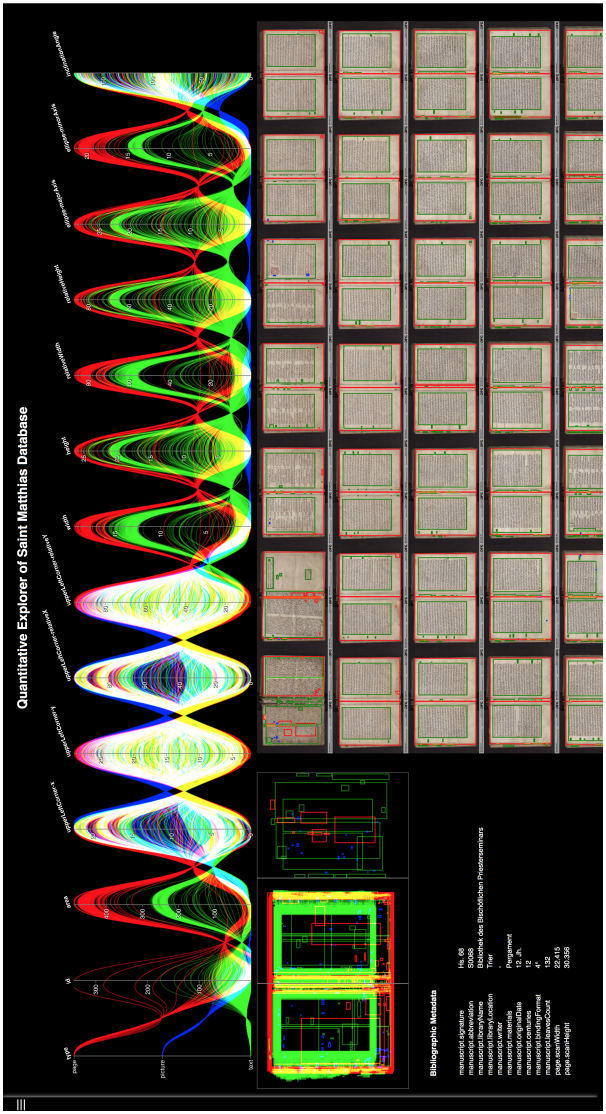


Figure 6: This figure shows a manuscript page explorer view. It shows all the measurements of layout features of a single manuscript. At the top, the parallel coordinate plot is shown where each line represents one measurement of a layout feature. At the bottom left, the superimposition plot is shown where all the measurements are superimposed to see the overall structure of hundreds of manuscript pages at a single glimpse. At the bottom right, the manuscript montage plot is shown where measurements are drawn on the respective manuscript page.

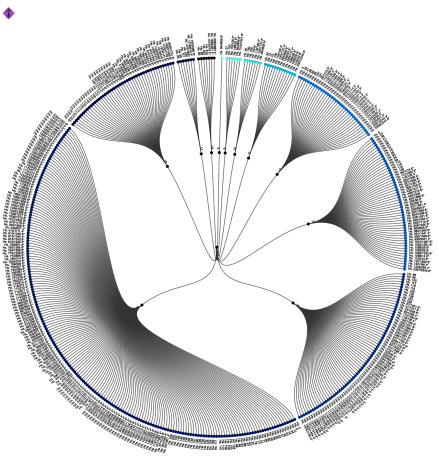


Figure 7: Radial tree with all 450 manuscripts in chronological order clockwise, starting with the 8th century. Multiple entries for manuscripts with more than one dating.

which is an attempt by the computer scientists to enable the (humanities) scholars to retrieve the information they want, make sense of that information, determine correlations in the data and reach decisions in a short period of time. For the prototype of CodiVis we have a radial tree (see fig. 7) providing an overview of bibliographical data, in this case the "century variable" of the manuscript according to the St. Matthias catalogues. The radial tree is a node-link tree with transformations in polar coordinates and was chosen over a simple chronological listing because it has a better usage of space if only few hierarchy levels and bottom nodes exist. To distinguish the different parameters and single manuscripts, different colours are used varying in brightness.

The layout features extracted by the SWATI workflow and other bibliographical metadata are represented using parallel coordinates (see fig. 8), a way of visualising high dimensional data and analysing multivariate data. The sample features of our prototype are number of columns, page height, page width, text hight and text width (mean value in cm), and text area (in cm^2). The layout features are mapped onto a vertical axis and each data value from the CSV file is represented along a line. It is scaled to lie between minimum and maximum at the top (see fig. 8). A pure collection of points would not be useful, so the points belonging to the same record are connected with lines. The colour assignment is similar to the radial tree. The arrangement of the vertical axis can be adjusted according to the correlations the

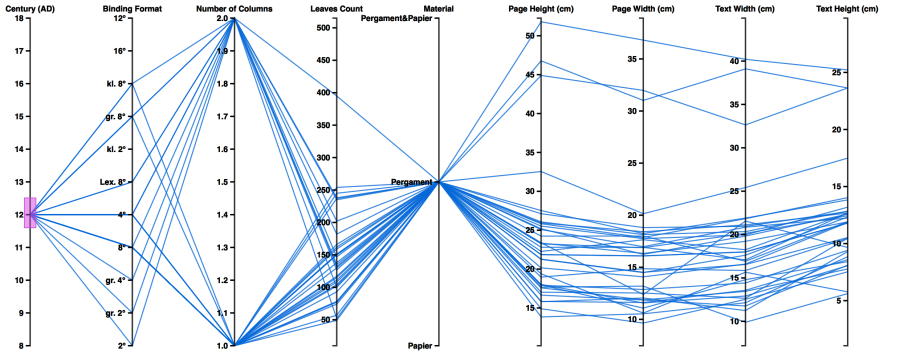


Figure 8: The image shows the parallel coordinate view with nine parameters showing manuscripts dating to the 11th century.

user wants to visualise. A table view where each row and column of the CSV file is represented completes the CodiVis framework. It is linked to both radial tree and parallel coordinate view. By brushing and linking, the static view becomes dynamic and interactive: two brushes are provided for analysing the mediaeval manuscript data. With the polar brush, users can select different nodes in the polar coordinates of the radial tree, the selected data are automatically reflected in the parallel coordinates and in the table view. With the vertical axis brush, users can brush any of the vertical axes of the parallel coordinates view. The selection in felected respectively in all three views.

Multi-dimensional visualisation techniques not only provide the humanities scholar more “beautiful data”, they can also help to explore single manuscripts and corpora. To prove the theories of notable rectangles, one just has to take a look at the vertical axes of the parallel coordinate view. Scholars accessing visualisations can easily choose a group of pages or manuscripts or even the whole library from the radial tree or *CodiStore* (see Chandna et al. 2016) database to generate a visualisation of the proportions of page and text. Two or more different layouts within one codicological unit can help to distinguish different parts of a composite manuscript. Outliers and pecularities within a group of manuscripts can be detected more easily.

Unfortunately, at the preliminary endpoint of the project we have not yet been able to experiment with our results and perform sufficient analyses to contribute results in quantitative layout studies. Processing big data such as our five terabyte of manuscript images is a complicated and time consuming task. Same applies to the selection of the right visualisation forms to facilitate access to data and to fulfil the

needs of different groups of researchers. The presented forms of visualisation are purely exploratory and have not been conclusively evaluated yet. Notwithstanding, we hope this refelction on the experiment *eCodicology* can show the potential of SWATI and CodiVis and we are looking forward to be able to give access to our developments, present further analyses, and discuss about the approaches of the project.[20]

Bibliography

Agati, Maria Luisa. *Il libro manoscritto. Da Oriente a Occidente. Per una codicologia comparata.* Rome: L'erma di Bretschneider (=Studi archeologica), 2009.

Becker, Petrus. *Die Benediktinerabtei St. Eucharius-St. Matthias von Trier.* (Germania Sacra, Neue Folge, Bd. 34) Berlin, New York (NY): de Gruyter, 1996.

Chandna, Swati, Danah Tonne, Thomas Jejkal et al. "Software workflow for the automatic tagging of mediaeval manuscript images (SWATI)." In Ringger, Eric K. and Bart Lamiroy. *Document Recognition and Retrieval XXII*, San Francisco, California, USA, February 11-12, 2015. SPIE Proceedings 9402, SPIE 2015.

Chandna, Swati, Danah Tonne, Rainer Stotzka et al. "An effective visualization technique for determining co-relations in high-dimensional mediaeval manuscript data." *Electronic Imaging, Visualization and Data Analyses* 2016. VDA-488/ 1-6.

Chandna, Swati, Francesca Rindone, Carsten Dachsbacher et al. "Quantitative exploration of large mediaeval manuscripts data for the codicological research." *Proceedings of IEEE Symposium on Large Data Analysis and Visualization* 2016. Baltimore (MD): IEEE, 2017. DOI: 10.1109/LDAV.2016.7874306.

CodiKos. <https://github.com/JochenGraf/CodiLab/blob/master/CodiKOS.html>.

eCodicology: *eCodicology – Algorithmen zum automatischen Tagging mittelalterlicher Handschriften.* Darmstadt: Technische Universität Darmstadt, Trier: Trier Center for Digital Humanities, Karlsruhe: Karlsruher Institut für Technologie. 2013-2016. <http://www.ecodicology.org>.

Heer, Jeffrey, Michael Bostock, and Vadim Ogievetsky. "A Tour through the Visualization Zoo. A survey of powerful visualization techniques, from the obvious to the obscure." *Communications of the ACM.* 53.6 (2010). 59-67.

ImageJ: *Image Processing and Analysis in Java.* <http://imagej.nih.gov/ij>.

Jakobi-Mirwald, Christine. "Lost in Translation. Manuscript terminology between languages." *Gazette du livre medieval* 55 (2009). 1-8.

Keuffer, Max and Gottfried Kentenich. *Beschreibendes Verzeichnis der Handschriften der Stadtbibliothek zu Trier.* Bd. 1—10. Trier: Lintz, 1888-1931.

KIT Data Manager. Karslruhe: Karlsruhe Institute of Technology. <http://datamanager.kit.edu/>.

[20] Updates about the availability of our software will be published on the *eCodicology* website. Don't hesite to contact the authors for further information or discussion.

Maniaci, Marilena. "Ricette di costruzione della pagine nei manoscritti greci e latini." *Scriptorium* XLIXC (1995). 16-41.

Maniaci, Marilena. "Ricette e canoni di impaginazione del libro medievale. Nuove osserazioni e verifiche." *Scrineum rivista* 10 (2013). 1-48.

Marx, Jacob. "Handschriftenverzeichnis der Seminar-Bibliothek zu Trier." *Trierer Archiv*. Ergänzungsheft 13 (1992).

MOA/WEKA: *Massive Online Analyses*. Version 16.04. April 2016. <http://moa.cms.waikato.ac.nz/>.

Montebaur, Josef. "Studien zur Geschichte der Bibliothek der Abtei St. Eucharius-Matthias zu Trier." *Römische Quartalschrift für Christliche Altertumskunde und Kirchengeschichte*. Supplementband 26. 1931.

Ornato, Ezio. "La codicologie quantitative, outil priviligie de l'histoire du livre medieval." *Historia, instituciones, documentos*. 18 (1991). 375-402.

Ornato, Ezio (ed.). *La face cachée du livre médiéval. L'histoire du livre vue par Ezio Ornato, ses amis et ses collègues*. Rom: Viella, 1997.

SemToNotes: *Semantic Topological Notes*. Cologne: University of Cologne, Institute of Humanities Computer Science, DARIAH-DE Phase II 2014-2015. <https://hkikoeln.github.io/SemToNotes/>.

TEI-C: TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Chapter: "Manuscript Description"*. Version 3.0.0. 2016. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>.

TextGrid: TextGrid Consortium. *Text Grid: A Virtual Research Enviroment for the Humanities*. Göttingen. 2006-2014. <http://textgrid.de>.

ROMA: *ROMA generating customazation for the TEI*. <http://www.tei-c.org/Roma>.

R-Project: *The R Project for Statistical Computing*. <https://www.r-project.org>.

Virtuelles Skriptorium: *Virtuelles Skriptorium St. Matthias*. Trier: Trier Center for Digital Humanities, StB/StA Trier. 2010-2014. <http://www.stmatthias.uni-trier.de>.

Image and Text in Numbers: Layout Analysis for Hispanic and Spanish Modern Magazines

Nanette Rißler-Pipka

Abstract

Hispanic and Spanish modern magazines were long time a neglected field of research. Even if the magazines are regarded as a valuable source for information about contemporary cultural, social and political life for various disciplines like cultural, literary, media or social studies or linguistics. But stored in libraries as sensitive material and threated soon by decay most of the magazines were not accessible for researchers. Since digitization this has changed. The project *Revistas culturales 2.0* (University of Augsburg) tries to work on the digital collection of the IAI (Ibero-American Institute, Preußischer Kulturbesitz, Berlin) by annotating the magazine pages and analysing the metadata using digital tools. But as we are experienced in cultural, media and romance studies, the complex field of automatic document analysis stayed hidden for us, without the cooperation with experts in computer science (where in context of digitizing projects and OCR important research is already achieved).[1] Also, in the DH this field is not as advanced as for example text mining, therefore it proved difficult to find tools for quantitative analysis (for example relation of textual and image parts in the magazines). In cooperation with the project *eCodicology* and reusing their tool SWATI we found now a way how layout analysis for Hispanic and Spanish modern magazines might be done in future by measuring each page automatically. In consequence the paper can only present a concept of how useful the tool and quantitative analysis of the layout might be for analysing the Hispanic and Spanish modern magazines from the perspective of Humanities.

Zusammenfassung

Lateinamerikanische und spanische Kulturzeitschriften der Moderne konnten lange nicht für Forschungszwecke im Original herangezogen werden. Dabei sind sie als wertvolle Quellen als Zeitzeugen des kulturellen, sozialen und politischen Lebens anerkannt und könnten in verschiedenen Disziplinen wie Kultur-, Literatur-, Medien- oder Sozialwissenschaften sowie in der Linguistik genutzt werden. Jedoch waren sie als empfindliches und leicht vergängliches Material in den Bibliotheken unter Verschluss und auch aufgrund geographischer Entfernung für viele Forscher

[1] See also the Pattern Recognition & Image Analysis Research Lab of the University of Salford Manchester (PrimA).

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 25–42.

unerreichbar. Seit der Digitalisierung hat sich dies geändert. Das Projekt *Revistas culturales 2.0* (Universität Augsburg) versucht nun mit der digitalen Sammlung des IAI (Ibero-Amerikanisches Institut, Preußischer Kulturbesitz, Berlin) zu arbeiten. Dazu werden Metadaten angereichert und vorhandene Metadaten mithilfe digitaler Tools analysiert. Als Geisteswissenschaftler musste uns bisher das Gebiet der automatischen Bildanalyse verschlossen bleiben, da dies vor allem auf technischer Seite in der Informatik (im Kontext von Digitalisierung und OCR) weiterentwickelt wurde.[2] Auch in den DH ist dieses Gebiet weit weniger bearbeitet worden als z.B. Text Mining und es erwies sich als schwierig, Tools für quantitative Analysen (um z.B. die Bild-Text-Relation in den Zeitschriften zu beziffern) zu finden. Nun konnten wir in Kooperation mit dem Projekt *eCodicology* und durch Nachnutzung ihres Tools SWATI ausprobieren, wie das Layout der lateinamerikanischen und spanischen Kulturzeitschriften der Moderne durch automatisches Vermessen jeder einzelnen Seite in Zukunft analysiert werden könnte. Der vorliegende Beitrag entwickelt aus einem ersten Experiment mit einer kleinen Anzahl von Zeitschriftenseiten ein theoretisches Konzept, um den Nutzen von quantitativen Methoden und im konkreten Fall vom angewendeten Tool aus Sicht der Geisteswissenschaften abzuschätzen.

1 Modern magazines as complex work of art

When libraries all over the world began to scan their treasures of thousands of pages of magazines, produced in the second half of 19th century until the second half of 20th century, many researchers (particularly in the Anglo-Saxon world) realized the potential that lies in the gained data. For the period of Modernity, the impact of cultural magazines on literary, artistic and social life is commonly known. The fast changing society required a medium that represented this acceleration in weekly or monthly issues. As well as the increasing linkage between fine arts, literature, photography, architecture, fashion, life-style, etc. was looking for a medium that incorporates the cultural change altogether. Furthermore, technical progress in printing industries and aesthetic ideas coming from the rising avant-garde culture produced an amazing number of new and partly ephemeral cultural magazines all over the western civilization (including Latin America). But to consider cultural magazines as an independent and complex work of art, which deserves to be analysed as a whole (i.e. each title of a magazine and the relation between magazines), has not been standard in literary, cultural, media or social studies for many years. Traditionally cultural magazines were used like a library of rare texts from known authors or as

[2] Ibid.

information reservoir for contemporary witness. The function of magazines was rather the one of an archive than of an original work of art worth to be looked at as a whole (Ehrlicher and Herzgsell 2016; Podewski 2016).

Now the digitization of numerous magazines initiated a change in their perception as complex cultural artefacts (Louis 2014; Pita González 2014; Maíz 2011; Stead and Védrine 2008). Equally the various changes in human perception in general which we can observe in Modernity are causally linked to the perception of magazines. People read the magazines in a public space and browse through them. They read them in fragments, are attracted by a picture or a headline and then interrupt their reading because of other distraction. For the cultural area of Latin America Raquel Macciuci draws the comparison between reading a magazine (or paper) and browsing the internet (Macciuci 2015, 219). Though, we have to take into consideration that contemporary readers and researchers reading and analysing the magazines today have a totally different kind of perception (Louis 2014, 33). In contrast to the contemporary reader we are able to have a look on all the issues of each magazine at once. This overview enables us to observe changes, particularly in the layout, but also regarding the staff and content of the magazine (the latter would be called metadata in the digital era). But to get an overview in a visual sense, we have to find a representation that allows us to look at many issues at once and to still recognize differences and changes in layout.

Before digitization, particularly in the Spanish speaking countries, most of the magazines were not accessible for researchers as they were stored in libraries in Latin America, Spain and elsewhere. Still, no researcher is able to read the massive number of pages produced in Spanish speaking magazines published in the period around 1850–1945. The research project *Revistas culturales 2.0* lists 585 titles of magazines, of which 224 titles are digitized, every title consists of 1–300 volumes with ca. 20–200 pages each. Directly accessible and annotatable via the website of our research project *Revistas culturales 2.0* are 23 titles of magazines, provided by our cooperation partner: Ibero-American Institute (IAI, Preußischer Kulturbesitz, Berlin). The small number of 23 titles still contains 477 single issues and about 23,000 pages. To read and analyse all of them is certainly possible, but time consuming and not very productive when the research question would be a comparison of layout, aesthetical and conceptual changes within and between the magazines.

For the image driven period of "modernism" the aesthetics are defined easily by layout: graphical elements, ornamental framings, printing types, etc. Therefore, the digital representation of the whole view of a magazine is important for projects like the *Modernist Journals Project* (MJP) or the *Modernist Magazines Project* (MMP) or the *Blue Mountain Project* (BMP). Nevertheless, none of these projects are able to do quantitative layout analysis for their document repositories.

2 How to analyse the layout of modern magazines using digital tools?

Like most of these Anglo-Saxon projects also the repository of the *Revistas culturales 2.0*-project consists of images, i.e. jpg-files. OCR is very complicated for the text written in columns and the mixture of text and image in the magazines. The *Blue Mountain Project* is the only one providing also the text of their rather small collection, but depending on the quality of the original, the OCR results might be poor.[3] For the few Spanish and Hispanic repositories of cultural magazines or periodicals providing also text gained by OCR, the plain text is not accessible and results for searching in the texts show poor OCR quality (Rißler-Pipka 2014, 60).

As layout and whole visual appearance of the modern magazines are very important, it is meaningful to analyse the layout of the magazines in their historic context – and not only to read the content. In the layout of the document the text is still quantitatively represented. The position and quantity of text is as important as the one of illustration, photograph, painting and other visual elements in the magazines. The problem is rather that digital research for document or image data is not in the same way advanced as text analysis (i.e. text mining, etc.). For users not experienced in computer science, it seems difficult to find a tool that is accessible and explained in a way other disciplines (like DH) except computer science are able to work with. In DH very celebrated, but also critically discussed tool ImagePlot by Lev Manovich and the Software Studies Initiative has certainly more potential than just plotting images into one, but it actually structures a great number of images on the basis of mostly two different features (represented in x- and y-axes). To be able to use the tool you need already measured and saved data behind the chosen features, for example average image saturation and average image brightness for each image in your collection (as in the Van Gogh example Manovich gives: 2015, 25–26). Manovich tested the tool also with the title pages of the *Times* by comparing the saturation of colour covers and brightness of black and white covers (Manovich and Douglas 2009). But for more complex data like the title pages of Hispanic modern magazines it comes already to its limits. Which kind of feature should be extracted to compare the magazines (or even only the title pages)? Brightness or colour saturation is not very useful here. When Manovich formulates questions for a collection of more than 10 million images like: "what are the subjects of these images" (Manovich 2015, 25), he is not going to answer them by using ImagePlot. For answering the question or even try to answer it, he uses primarily metadata (in the case of *On Broadway* he is not answering the question

[3] See the commentary in the Blue Mountain Archive (BMP): "Issues with poor quality paper, small print, mixed fonts, multiple column layouts, or damaged pages may have poor OCR accuracy. The searchable text and titles in this collection have been automatically generated using OCR software. They have not been manually reviewed or corrected."

Figure 1: Title pages of 18 magazines (1898–1931) with 245 issues from Latin America (IAI Collection)

by computing at all, but already the metadata provided by the source of the image (Instagram and social media statistics, Taxi statistics, Google Street View, etc.) gives information about the subjects of the images). In the context of document analysis as a field of computer science this approach is not even mentioned (Doermann and Tombre 2014).

Nevertheless, trying to look at the corpus in a bird's eye view (fig. 1), we can ask: What do we actually see, when looking at the miniatures of title pages? We are able to detect colour vs. black/white printing and we roughly detect images or bigger headlines vs. mostly text. If we would add more title pages in one image, we would see no details at all. In this case you need feature extraction and some computing before plotting them together. The difficulty still is, which are the features worth to extract and to compare? When Manovich claims that, "Computer can identify regions that have similar colour value and measure orientations of lines and properties of

texture in many parts of an image" (Manovich 2015, 22), he is not able to explain how it works. Lately Waltraud von Pippich proved the difficulties analysing paintings by using the computer. She points out that Manovich's analysis depends on pixel resolution and hardware. Plus, the comparison of mean values cannot represent the features of an image: "Zur Extraktion farbformaler Bildeigenschaften ist diese Methode der Medienkunst ungeeignet" (Pippich 2016).

For humanists the challenge is to see the image from a computer's perspective, that means without any semantics. For the computer text or image is just a different distribution of colour, brightness and other features. To find out, what could be the interesting features to be automatically extracted, we tried first a traditional interpretation of layout, keeping in mind the perspective of the computer.

2.1 Stepping back: Examples for traditional layout analysis

By zooming in the title pages (fig. 2) we could see changes in the development of some magazines, while others stuck to their layout for the whole period of their lifetime.

The Peruvian magazine *Amauta* has one of the most colourful and interesting layout in the collection (plus it is commonly regarded as one of the most important avant-garde magazine in Latin America). Initially repeating the typified Inca-head on the cover and then experimenting with various ideas of recognizable images and symbols until the last number in 1931 repeats again the initial Inca-head (appearance of the head 7 times over the years, for 31 issues in all). Other recurring elements are the rising sun (no. 10 and 30), the sowing Farmer-Inca (3 times: no. 12–14), the stairs (no. 19, 23, 28), the mixture of writing and mask in two different versions (no. 20 + 27 and 26 + 29) and finally the writing in combination with the initial Inca-head (no. 22, 25).

But working only with title pages can be heavily misleading. In the case of *Amauta* the fascinating aesthetics represented by the changing cover truly reflects the status of an avant-garde-magazine, but the inside of the magazine is not at all colourful or adventurous regarding the layout (fig. 2).

Typeface, paragraphs and the position of the images are not at all revolutionist or courageous, but rather traditional. The segmentation of the pages in text, image, headlines, paragraphs, even the decorative capital letter at the beginning of each article, poem, etc. can be described as conventional layout. Only the recurring ornaments or hieroglyphics of the Inca-culture are slightly unusual. Nevertheless, in the combination of the intriguing and consistent indigene aesthetics and the traditional layout it is recognized as a celebrated avant-garde magazine for Latin America. The obvious emphasis on the indigene roots of Peru fits to the fascination for indigene cultures in general in the international avant-garde movements. The editor Mariátegui brings his impressions of recent journeys to Europe into the concept of the magazine:

Figure 2: The title pages of *Amauta* (1926–1931)

"Yo vine de Europa con el propósito de fundar una revista" (Maríategui 1926, 1).[4] He tries to combine avant-garde-ideas and socialism as well as pro-Indio and European ideas. Taking into consideration the rich background information about Maríategui and his engagement in magazines (Beigel 2006; Manzoni 2004; Melgar Bao 2006), which can also be called metadata in the DH-sense, we have to evaluate the layout of *Amauta* in a different way.

The courage in layout and aesthetics of *Amauta* only shows when comparing it to other magazines of the same period, place and cultural context. The evaluation of the composition of image and text in the magazine should consider the contemporary standard.

[4] Translation: "Coming back from Europe, I had the idea of funding a magazine".

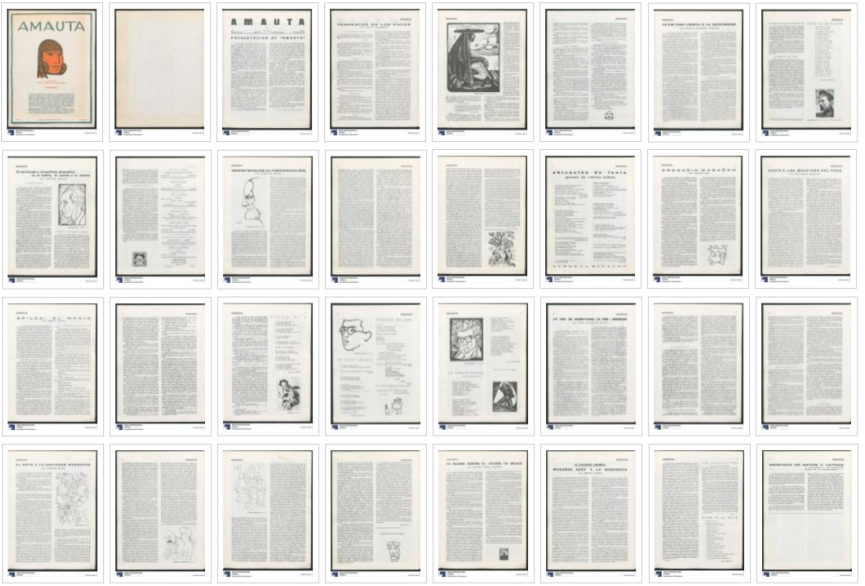


Figure 3: The first issue of *Amauta* (No. 1, 1926)

The rather conservative and elitist magazine *La Nueva Revista Peruana* shows in contrast to *Amauta* no images at all (fig. 4). This layout can be called book-look-alike and has certainly no interest to attract the reader's attention by visual aspects. Even if the editorial speaks of "una visión sin compromisos", this sentence is continued by looking back to the old values: "aunque transfigurada por el fuego de un antiguo fervor" (Editorial, Nueva Revista Peruana 1929, No. 1, p. 2).[5]

Even earlier, but representing another subgenre of cultural magazines, the *Ilustración Peruana* (1911–1912) is rather made to be browsed through and for distraction than for reading intellectual and cultural debates (fig. 5).

Yet here the layout can be misleading, because the distraction and rather decorative presentation of photographs, illustration and less text does not fit to the philosophical character of the first article "Nuestros Problemas – Valor y Trabajo" by Don Alejandro O. Escarza. Still conservative, the author requests some thinking from his reader by discussing on three long pages the problem of education in Peru. In contrast to Maríategui and his magazine *Amauta*, Escarza represents the conservative elite of Peru who are not willing to 'waste' money for the education of Indios.

[5] Translation: "a vision no-holds-barred … but expressed with the fire of former intensity".

Figure 4: *La Nueva Revista Peruana* (1929, No. 1)

These three Peruvian magazines show very clearly that neither the layout nor the metadata or textual content alone can be sufficient for an analysis. Plus, the variety concerning layout and concept of cultural magazines in such a short period of 1912–1929 in Peru is obviously quite large.

2.2 The other way – or why do we need digital tools?

What have we learned trying to analyse the three Peruvian examples of modern cultural magazines? We looked at 3 titles of magazines, part of 3 issues and at 72 pages and can say without looking at the rest of the pages of the three titles that each of them has a recognizable and individual layout. You could probably allocate correctly every single page of the magazines to the right title without reading a sentence of them. Though, allocating them does not mean knowing anything about them, apart from knowing that they belong together and that they are part of a magazine.

That means layout can be one criteria to distinguish different titles of magazines in different cultural contexts and different subgenres. But layout does not necessarily correspond to other semantics and can easily be misleading. To draw the deduction that more illustrations means necessarily a more popular magazine can be as false as

Figure 5: *Ilustración Peruana* (No. 152, 1912)

the deduction that a traditional layout only fits to a traditional magazine. Still, the whole visual appearance of a magazine is certainly part of its aesthetics and concept (in a cultural, social, political, artistic sense). If the content and other elements really fit to the visual signals is another question.

The variety in layout and other visual elements of modern magazines is rather numerous as we have seen in the three examples above. For an analysis of layout in a quantitative way the possibility to deduct general hypothesis on the basis of some exemplary analysis seems to be difficult and speculative. 'Counting' the layout as an interplay of text and image in numbers promises a more reliable method. The chance to analyse all of the ca. 23,000 pages in the IA-collection – and even more, if other repositories are used – is worth the effort of learning to handle digital tools and to ask the right questions to know which features of the document should be extracted and how to analyse the results. The example of Lev Manovich shows how difficult it is to find the right tool and that computer science is needed to understand the functionality and perspective of automatic document analysis. Therefore, we were happy to find in the cooperation with the project *eCodicology* not only the fitting tool, but also the explication of it (Chandna et al. 2016; Chandna et al. 2015). Even if the handling of

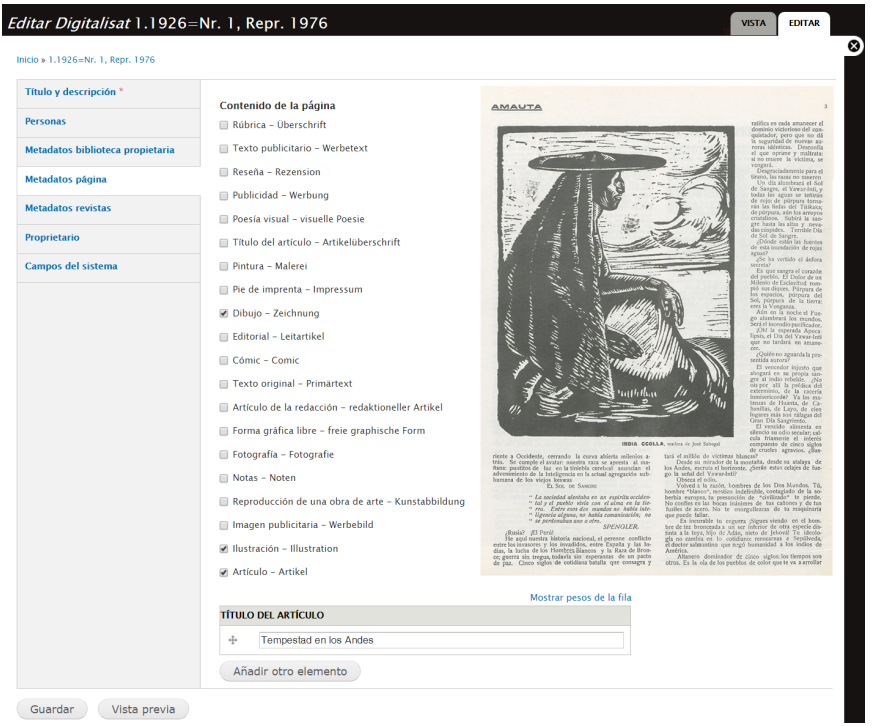


Figure 6: User interface for annotation, gathering metadata, example, *Amauta* (No. 3, 1926)

the tool had to be left to the experts (KIT, Karlsruhe Technology Institute) and we can only interpret the results, but not the technical function in detail. In this point further cooperation is needed.

Nevertheless, the manual or traditional layout analysis of the three Peruvian examples also showed, that we need more metadata to draw conclusions which are reliable and based on more information than just the visual one. On the project website of *Revistas culturales 2.0* we implemented the possibility to annotate each page of the collection (fig. 6).

While gathering metadata on the basis of crowd-sourcing the problem is that not enough people are willing to do the annotating of Hispanic modern magazines, because the community is not yet used to work with digital corpora. However, the metadata we already get by the IAI (title, year, contributors, place, country) and the one gathered manually by project collaborators help to do geographic overviews (see fig. 7) or network visualizations (Ehrlicher and Herzgsell 2016).

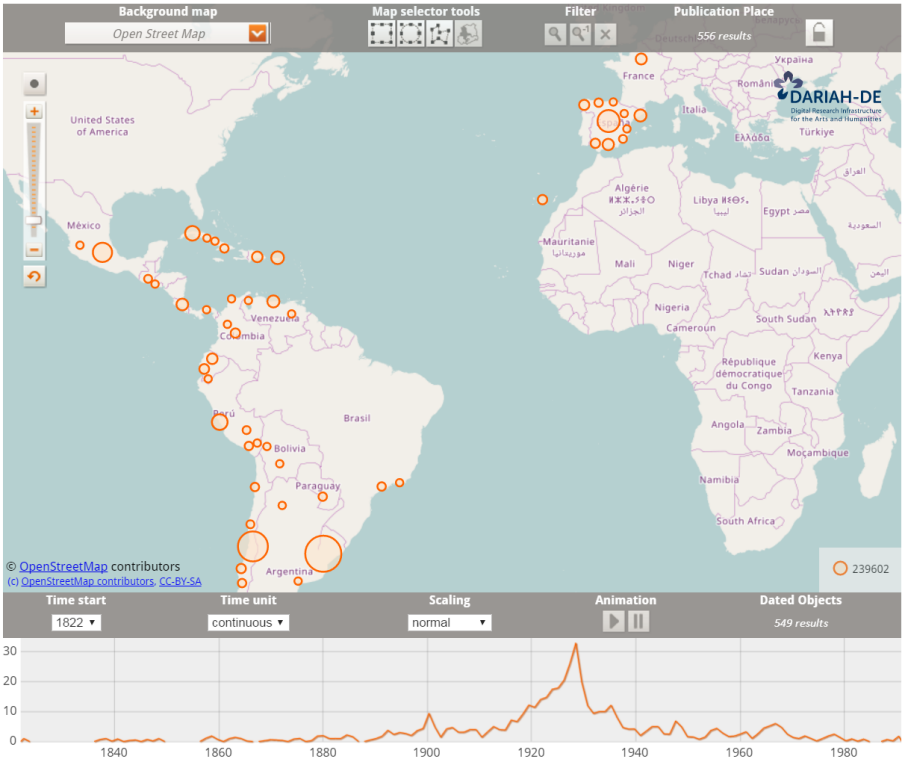


Figure 7: The bibliography of Hispanic and Spanish modern magazines in *Revistas culturales 2.0* visualized by DARIAH Geo-Browser

This example also shows why digital tools may help to analyse modern magazines in their cultural context. The distribution of magazines (as visible in the timeline, most of them appeared 1900–1940, with a peak in the 20s) and the spreading all over the Spanish-speaking world is amazing. The metadata is structured and ready for interactive use (for example to see all titles published in a chosen place or at a chosen time). Still, structured metadata alone does not help to know anything about the visual appearance of each magazine.

3 The experiment: Trying a tool for medieval manuscript analysis

The reuse of existing tools in a completely different context is rare, because every project designed their tool for exactly the data and research questions they may have. But we tried, if a tool designed for the feature extraction in medieval manuscripts can be adopted for measuring also pages of cultural magazines. From the computer science point of view this is a complex and difficult process, which is not transparent for me as a researcher in humanities. Nevertheless, knowing that it is possible enlarges the possibilities for cultural analysis. Together we think about the question: which features should be extracted? But as we already knew the features that 'can' be extracted by using the tool SWATI, we now have an idea of how it works:

"Layout features of the medieval manuscripts extracted by SWATI: Number of Pages, Mean Colour Value, Page Width, Page Height, Upper Left Corner Coordinates of Page, Relative Measurements of the Page, Text Width, Text Height, Text Areas, Upper Left Corner Coordinates of Text, Relative Measurements of the Text, Pictorial Width, Pictorial Height, Number of Pictorial Areas, Upper Left Corner Coordinates of Pictures, Relative Measurements of the Pictures" (Chandna et al. 2016, 3).

The principle differentiation we are interested in, is the one between text and image. In general, we know how important this difference is for analysing magazines – even if the quantitative distribution of image and text can be misleading as we have seen in the example of *Ilustración Peruana* (see fig. 5). By the named features SWATI is extracting we would also know how many text- and picture-areas we have on each page and additionally which dimensions they have and which position in the page. Certainly there are more complex questions to be asked regarding modern magazines like the graphic elements, ornaments and printing types, but for the beginning the tool would do more than expected.

So, we transferred some examples of magazine pages provided by the IAI and the metadata for the scanning process (colour checker, etc.) to the KIT, where Swati Chandna tried if the tool is working with this different kind of document. Thanks to her (and the whole *eCodicology* team) we got the measurements for image and text separately for six example pages of the magazine *El Hogar* (Dec. 1919). The results are visible in "Image" and "Numbers" (fig. 8-10).

Comparing the image and text segmentation in the examples we see how the tool is working. Particularly for Humanists the direct comparison between original and image/text segmentation is very useful to understand the technical and mathematical process behind. Even the quite complicated mixture of image and text in the advertisements is correctly recognized, if you regard the graphical elements in the headlines of the advertisements as image. Indeed, they function as both: image and text, because they should attract the reader's attention by visual appearance and give

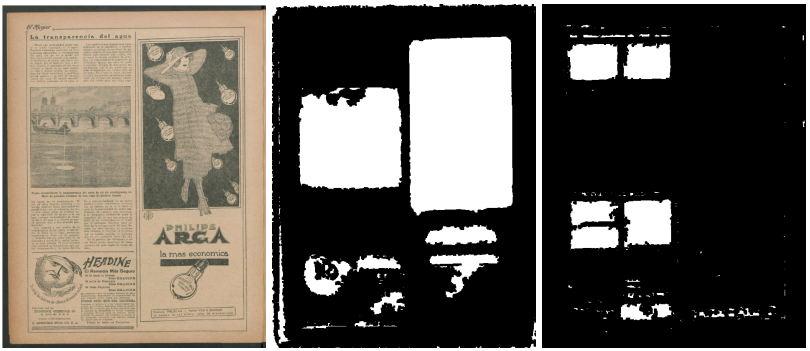


Figure 8: Original page – image segmentation – text segmentation

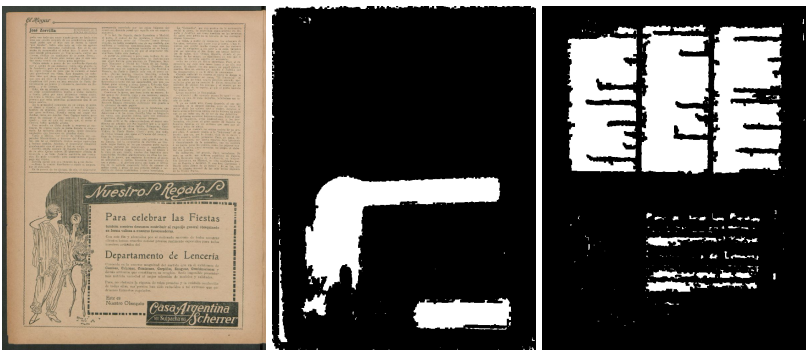


Figure 9: Original page – image segmentation – text segmentation

additionally textual information. The only problem, that might be resulting out of the measurement is the recognition of textual and pictorial units. For example, the advertisement in figure 9 is divided in a pictorial part and a textual part. In the tables of measurements, you won't know if these parts belong together or not – apart from using the extracted data for the position on the page and drawing the conclusion that textual and pictorial part belong together if they are positioned that near.

We hoped the results in numbers would be something like: For the title xy we count xxx images and xxx texts, so that we could compare these numbers with other results for other magazines, periods, places, editors, etc. But we had to learn, that measuring pages automatically is far more complicated. For both sides, the technical one and the humanist one, it is important to see by experimenting which limitation there are and to work together on further research to widen the possibilities.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Signature	Type	ID	Area	Mean	Min	Max	BX	BY	Width	Height	Major	Minor	Angle
2	6	page	1	885.241	255	255	255	0.517	0.475	26.314	33.873	38.046	29.625	89.963
3	6	picture	2	1.303	255	255	255	0	0	5.610	1.025	3.915	0.424	3.447
4	6	picture	3	83.635	255	255	255	3.356	17.280	19.831	11.763	16.629	6.404	22.026
5	6	picture	4	2.960	255	255	255	0.559	24.924	1.017	9.297	8.229	0.458	91.779
6	6	picture	5	3.200	255	255	255	2.975	29.254	2.551	2.254	2.268	1.797	12.263
7	6	picture	6	23.637	255	255	255	14.356	29.720	9.975	3.144	10.708	2.811	2.057
8	6	text	7	1.247	255	255	255	13.288	1.915	3.356	0.678	3.333	0.477	2.339
9	6	text	8	83.474	255	255	255	10.110	2.000	6.831	15.068	15.513	6.851	90.172
10	6	text	9	1.039	255	255	255	7.347	2.085	2.297	0.763	2.037	0.649	178.675
11	6	text	10	81.214	255	255	255	17.364	2.085	7.093	14.983	15.310	6.754	88.771
12	6	text	11	79.294	255	255	255	2.593	2.847	7.051	14.093	14.523	6.952	88.767
13	6	text	12	2.814	255	255	255	10.576	23.441	5.856	1.017	4.793	0.747	3.170
14	6	text	13	1.840	255	255	255	10.958	26.881	4.542	0.890	4.009	0.584	179.296
15	6	text	14	3.766	255	255	255	10.534	27.941	4.966	1.186	4.947	0.969	179.561

Figure 10: Table of measurements for textual and image parts, for one page – only biggest values considered (*El Hogar*, Dec. 1919)

For now, the measurements give us many numbers and even more images (see fig. 8-9, plus the page segmentation not illustrated here). That means on the one hand we reduced the complexity of each page, by focusing on pictorial and textual feature extraction, but on the other hand we produced much more complexity by all the data (in numbers and images) we gathered. All in all, the metadata gathered automatically is a result that cannot be valued highly enough. It is reliable because it is exactly the same process for each page, but should be observed and interpreted together with the people who programmed the tool.

What we tried here for some example pages should be easily done for all of the 23,000 pages of the IAI collection, but what we need now is a reduction of complexity for the gained data to be able to analyse it. That means we have to learn how to read and interpret the figures in the table (fig. 10). A first step to work with the given tables of measurement could be, to take only the biggest values for text and image, because all the tiny parts probably belong to one of the bigger ones. But as Swati Chandna points out, this won't work as a rule and it is necessary to observe the data first very carefully before defining the filtering rules.

Still, we don't know really how many entire images and texts are on the page. So, the next step would be to do statistical analysis and visualization of the metadata. As described for the visualization tool CodiVis in the *eCodicology* project this might be also possible for the modern magazines (Chandna et al. 2016, 3–5). Another perspective is, that the few examples show already that the distribution of textual and image parts on each page might give information of the kind of image or text, we look at (without really need to look at each page). The advertisements in *El Hogar* in all examples combine few textual parts (not forming an associated block) with image parts (also not forming the quadrat in which real illustration are usually represented).

Based on this observation we can try to build a model for recognizing advertisement semi-automatically. It might also happen, that advertisement for other titles than *El Hogar* use different forms of representation which can be described as different mathematic models to refine the tool with. Another challenge will be to recognize automatically the correlation between textual parts and image forming one ensemble. For example, in figure 8 we see that the image and the text below belong together and form one single advertisement, but in the automatic recognition of text and image parts this ensemble is correctly recognized as separate parts. The same is true for the example in figure 9.

The short experiment shows the different approaches in computer science (or at least DH) and Humanities. The measurements do not necessarily fit to the research question formulated in Humanities. That means more cooperation and exchange of ideas is necessary. But, looking at the results from a humanist point of view the effect is an estrangement (as Stephen Ramsay points out in *Algorithmic Criticism*). That means, we can now explore and observe aspects which were hidden for us before. For example, the fact that bold letters (in advertisements or other contexts) are recognized as image parts is meaningful also in aesthetics, because the function of these letters is at the same time a pictorial and writing one. The obvious failure of recognition becomes a double insight when observed in the humanist context.

4 Combined methods and knowledge

For a conclusion of the starting experiment, we can state that neither the traditional method of layout and magazine analysis can provide satisfying results for the whole corpus, nor the quantitative method only tested for now with some example pages can do the analysis of the magazines in their cultural context.

But a combination of the two methods could be a step towards an analysis of (Hispanic and Spanish) modern magazines considered as a complex interplay of text, image, content, political, social and cultural context. Problems to be solved in the future are:

1. Try the tool SWATI for the rest of the corpus, then try CodiVis and other statistics on the gained data.
2. Integrate the metadata of the magazines already gathered by annotating and the bibliography into the visualization for having more relation between the raw measurement and the context of each magazine. Therefore, we would be able to answer the difficult questions of an historic process: How much changed the appearance of cultural modern magazines from 1850–1945? What are the parameters of change?

3. Try a quantitative analysis and comparing this to the results gained already in secondary literature on Spanish and Hispanic modern magazines and in own exemplary analysis.

Acknowledgements

This work is the result of a cooperation between the project *Revistas-culturales 2.0* (University of Augsburg), the IAI (Ibero-American Institute, Preußischer Kulturbesitz, Berlin) and the project *eCodicology* (Technische Universität Darmstadt, Karlsruhe Institute of Technology, University of Trier). I thank the IAI for providing the data and the team of *eCodicology* for adapting and applying the tool SWATI on the data. For personal support and patient explication, I thank particularly Swati Chandna (KIT).

Bibliography

Beigel, Fernanda. *La epopeya de una generación y una revista: las redes editoriales de José Carlos Mariátegui en América Latina.* Buenos Aires: Editorial Biblos, 2006.

BMP: *Blue Mountain Project.* <http://bluemountain.princeton.edu>.

Chandna, Swati, Danah Tonne, Thomas Jejkal et al. "Software Workflow for the Automatic Tagging of Medieval Manuscript Images (SWATI)." In Ringger, Eric K., and Bart Lamiroy (eds.). *Document Recognition and Retrieval XXII.* 940206. 2015. DOI: 10.1117/12.2076124.

Chandna, Swati, Danah Tonne, Rainer Stotzka et al. "An Effective Visualization Technique for Determining Co-Relations in High-Dimensional Medieval Manuscripts Data." *Visualization and Data Analysis 2016*, San Francisco, California, USA, February 14-18, 2016. 1–6. <http://ist.publisher.ingentaconnect.com/contentone/ist/ei/2016/00002016/00000001/art00013>.

Doermann, David and Karl Tombre. *Handbook of Document Image Processing and Recognition.* London: Springer, 2014.

Ehrlicher, Hanno, and Teresa Herzgsell. "Zeitschriften Als Netzwerke Und Ihre Digitale Visualisierung. Grundlegende Methodologische Überlegungen Und Erste Anwendungsbeispiele." *Revistas Culturales 2.0*, 2016. <http://www.revistas-culturales.de/de/buchseite/hanno-ehrlicher-teresa-herzgsell-zeitschriften-als-netzwerke-und-ihre-digitale>.

Louis, Annick. "Las Revistas Literarias Como Objeto de Estudio." In Ehrlicher, Hanno and Nanette Rißler-Pipka (eds.). *Almacenes de un tiempo en fuga: Revistas culturales en la modernidad hispánica.* Aachen: Shaker, 2014. 31–57.

Macciuci, Raquel. "Técnica, soporte, ámbitos de sociabilidad y mecanismos de legitimación: sobre la construcción de espacios de literatura en la prensa periódica." In Schlünder, Susanne, and Raquel Macciuci (eds.). *Literatura y técnica: derivas ficcionales y materiales. Libros, escritores, textos, frente a la máquina y la ciencia. Actas del VII Congreso Orbis Tertius.* La Plata: Ediciones del lado de acá, 2015. 205-231.

Maíz, Claudio. "Las Re (D) Vistas Latinoamericanas Y Las Tramas Culturales: Redes de Difusión En El Romanticismo Y El Modernismo." *Cuadernos Del CILHA* 12.14 (2011). 75–91.

Manovich, Lev. "Data science and digital art history." *International Journal for Digital Art History* 1 (2015). 14-35. <https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/21631>.

Manovich, Lev and Jeremy Douglas. "Timeline: 4535 Time magazine covers, 1923-2009." 2009. <https://www.flickr.com/photos/culturevis/3951496507/in/set-72157624959121129.

Manzoni, Celina. "Las Formas de Lo Nuevo En El Ensayo: Revista de Avance Y Amauta." *Revista Iberoamericana* 70.208 (2004). 735–747.

Maríategui, José Carlos. "Presentación de *Amauta*." *Amauta* 1 (1926). 1.

Melgar Bao, Ricardo. "Mariátegui Y La Revista Amauta En 1927: Redes, Accidentes Y Deslindes." *Revista de Antropología* 4 (2006).

MJP: *Modernist Journals Project.* <http://www.modjourn.org>.

MMP: *Modernist Magazines Project.* <http://modernistmagazines.com>.

Pippich, Waltraud von. "Rot Rechnen." *Zeitschrift Für Digitale Geisteswissenschaften.* Sonderband 1: Grenzen und Möglichkeiten der Digital Humanities. 2016. DOI: 10.17175/sb001_016.

Pita González, Alexandra. "Las Revistas Culturales Como Soportes Materiales, Prácticas Sociales Y Espacios de Sociabilidad." In Ehrlicher, Hanno and Nanette Rißler-Pipka (eds.). *Almacenes de un tiempo en fuga: Revistas culturales en la modernidad hispánica.* Aachen: Shaker, 2014. 227–45.

Podewski, Madleen. "Mediengesteuerte Wandlungsprozesse. Zum Verhältnis zwischen Text und Bild in illustrierten Zeitschriften der Jahrhundertmitte." In Mellmann, Katja and Jesko Reiling (eds.). *Vergessene Konstellationen literarischer Öffentlichkeit zwischen 1840 und 1885.* Berlin: de Gruyter, 2016. 61–79.

PrimA: *Pattern Recognition & Image Analysis Research Lab.* University of Salford Manchester. <http://www.primaresearch.org>.

Ramsay, Stephen. *Reading Machines. Toward an Algorithmic Criticism.* Urbana: University Press, 2011.

Rißler-Pipka, Nanette. "Sobre Los Problemas de Investigación Con Revistas Culturales Digitalizadas Del Mundo Hispanohablante." In Rißler-Pipka, Nanette and Hanno Ehrlicher (eds.). *Almacenes de un tiempo en fuga: Revistas culturales en la modernidad hispánica.* Aachen: Shaker, 2014. 59–80.

Stead, Évangélia and Hélène Védrine (eds.). *L'Europe des revues (1880 - 1920): estampes, photographies, illustrations. Histoire de l'imprimé.* Paris: Presses de l'Univ. Paris-Sorbonne, 2008.

Revistas culturales 2.0: *Virtuelle Forschungsumgebung zur Erforschung spanischsprachiger Kulturzeitschriften der Moderne.* Universität Augsburg. 2014-2016. <http://www.revistas-culturales.de>

Bibliotheken im Buch: Die Erschließung von privaten Büchersammlungen der Frühneuzeit über Auktionskataloge[*]

Hartmut Beyer, Jörn Münkner, Katrin Schmidt, Timo Steyer

Zusammenfassung

Der Beitrag demonstriert anhand eines Auktionskatalogs von 1670 unser Vorgehen, frühneuzeitliche Gelehrtenbibliotheken bibliographisch nachhaltig zu erschließen. In einem ersten Schritt beschreiben wir die Erfassung der im Katalog verzeichneten Titel. Das Instrument für diesen Arbeitsgang ist eine Excel-Tabelle, die bibliographische Ermittlung erfolgt mit Hilfe nationaler und internationaler Online-Kataloge. Im zweiten Schritt geht es um die Entwicklung der digitalen Infrastruktur für die Onlinepräsentation der Daten. Hierzu wurde ein frei nachnutzbares Programm entwickelt, das für die Rekonstruktion frühneuzeitlicher Privatbibliotheken optimiert ist. Vorgestellt werden die verschiedenen textlichen und graphischen Visualisierungsformen sowie die weitergehenden Einsatzmöglichkeiten als Darstellungs- und Normierungstool für bibliographische Daten. Im dritten Schritt skizzieren wir den absolvierten Workflow und zeigen, wie traditionelle Methoden der Altbestandserschließung mit Verfahren der Digital Humanities kombiniert werden können. Dabei rückt auch die digitale Edition eines Briefwechsels in den Blick, der den Auktionskatalog als Sekundärquelle flankiert.

Abstract

Focusing on the 1670 auction catalogue of the books owned by a German expatriate living in the Netherlands, our article demonstrates a modus operandi for collecting bibliographical data and reconstructing private libraries from the early modern period. To begin with, we describe a method for the retrieval of title lots from the catalogue. The exact bibliographical data is supplemented by information from various national and international online-catalogues and databases. In a second step, we discuss the digital infrastructure for presenting the data on the web. For this purpose a multifunctional software for the reconstruction of early modern private libraries has been developed. Thirdly, we delineate the workflow and show how traditional

[*] Der Artikel ist im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekts *Autorenbibliotheken* im *Forschungsverbund Marbach Weimar Wolfenbüttel MWW* entstanden. Die vier Beitragenden geben entsprechend ihrer Zuständigkeit im Projekt Auskunft.

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 43–70.

methods of developing and providing access to library inventories can be combined with DH methods. Pertaining to this latter point, the article also discusses the digital edition of a correspondence between the expatriate and Duke August of Brunswick complementing the auction catalogue.

1 Private Bibliotheken und Auktionskataloge

Die mittelalterliche Mündlichkeits- und Manuskriptkultur veränderte sich mit der Etablierung der Druckerpresse im frühneuzeitlichen Europa immens. Gutenbergs ›Schönschreibhandwerk‹ mit beweglichen und wiederverwendbaren Lettern trieb die Ablösung der Kommunikation vom Körper voran, dehnte begrenzte Kommunikationsräume aus, inaugurierte eine neue Form von Öffentlichkeit und definierte Standardsprachen, aus denen sich die National- und Fachsprachen entwickelten (Coy 1994, 70). Produzierten die mittelalterlichen Schreibstuben Unikate oder wenige handschriftliche Kopien, die auch einzelne private Büchersammlungen bestückten, so sorgte der maschinelle Typendruck im Verbund mit den graphischen Reproduktionstechniken für hohe Ausstoßraten, eine potenzierte Distribution und Verfügbarkeit von Text-Bild-Kombinationen im öffentlichen und privaten Bereich. Mit der Akkumulation und Proliferation von Titeln zirkulierten Lektüren und Informationen, die debattiert, bestätigt und verworfen wurden. Wissen wurde zunehmend öffentlichkeitswirksam verhandelt und neu generiert. Dieser Wandlungsprozess, nunmehr im Zeitalter elektronischer Datenverarbeitung, dauert an. Die vor 40 Jahren einsetzende computertechnische Prozessierung der ehemals distinkten Medienformate Bild, Schrift, Zahl und mittlerweile auch Ton bedeutet eine nächste Zäsur in der Medienentwicklung und Zeichengeschichte. Freilich sind die alten Medienformate und Konstellationen damit nicht auf einmal obsolet, sondern es koexistieren wie in jeder Umbruchsituation die alten und neuen Medien. Zutreffend scheint jedenfalls, dass »die Gutenbergsche Galaxis der statischen Druckmedien […] in der Turingschen Galaxis der dynamischen programmierbaren Medien auf[geht]« (Coy 1994, 71).

In der Frühneuzeit, die den Horizont für die folgenden bibliotheksinvestigativen und digitalisierungspragmatischen Ausführungen stellt, also *grosso modo* zwischen dem 16. und 18. Jahrhundert, bekamen Privatpersonen erweiterte Möglichkeiten, eigene Bibliotheken aufzubauen. Diese erzählen immer mehr als die Bücher, die sie enthalten, zum Beispiel in welchen gelehrten Austauschbeziehungen Bibliotheksbesitzer standen, welche Lektürevorlieben gepflegt wurden, und anderes mehr. Nach dem Ableben der Besitzer wurden die Sammlungen indessen oft veräußert und unwiederbringlich auseinandergerissen. Erhalten haben sich zahlreiche Bestandsverzeichnisse, darunter viele Auktionskataloge, die Büchersammlungen für eine potentielle Käuferschaft und sonstige Empfänger repräsentieren, indem sie Titellisten präsentieren. Historische Verkaufskataloge, und von ihnen

sprechen wir hier ausschließlich, gleichen Fernrohren, die Buchbestände vor das Auge ziehen, deren physische Existenz zum Teil weit zurückliegt. Die Kataloge lassen die Profile von einstigen Sammlungen erkennbar werden, sie sind die Quellen, die begründete Vermutungen über die Lese- und Forschungsinteressen, die Arbeitsweisen, Leidenschaften und möglichen Netzwerke der ehemaligen Bibliotheksbesitzer zulassen. Des Weiteren können dank der Kataloge Verkaufs- und Preisroutinen für das Objekt Buch rekonstruiert werden (Cruz 2009, 105ff.; Hakelberg 2015a, 216f.; Raabe 1984, 277-280; Pozzo 2013, 8f.; Adam 2015, 69-72). So gelten Verkaufskataloge zurecht als eine »Hauptquelle des Buchbesitzes in der Frühen Neuzeit« (Ball 2008, 193). Jedoch ist bei der Beurteilung des Materials quellenkritische Vorsicht geboten, denn die formalisierten Sachtexte registrieren Momentaufnahmen von Bibliotheken; sie dokumentieren zwar den in einem bestimmten Augenblick vorliegenden Aggregationszustand einer Büchersammlung, geben aber kaum Aufschluss über ihre dynamische Veränderung. Daneben ist es nicht ausgeschlossen, dass Auktionatoren die zu versteigernden Sammlungen manipulierten, indem sie Titel entnahmen und Bücher fremder Provenienz hinzufügten. Grundsätzlich ist zu berücksichtigen, dass eine durch den Auktionskatalog (quasi) erschließbare Bibliothek »nicht einfach die inneren Denkvorgänge nach aussen [sic]« projiziert, sie also »kein ausgelagertes cerebrales Repositorium« per se darstellt (Wieland 2010, 28). Dieser Umstand relativiert den postulierten Bedeutungsgrad von privaten Büchersammlungen als erstklassigen Auskunftsquellen von Gedankenwelten, Ideenfindung, Weltbildern, Haltungen und Arbeitsstrukturen historischer Besitzer, Leser und Nutzer, zumindest wenn sie durch Auktionskataloge erschlossen werden.

Die Herzog August Bibliothek Wolfenbüttel (HAB) bewahrt mehr als 1700 gedruckte Buchauktionskataloge, deren Erscheinungszeitraum vom letzten Viertel des 16. Jahrhunderts bis in die Gegenwart reicht, darunter zahlreiche Unika. Ein statistischer Durchlauf registriert für das 16. Jahrhundert zwei Exemplare, für das 17. Jahrhundert 491, für das 18. Jahrhundert 543, für das 19. Jahrhundert 619 und für das 20./21. Jahrhundert immerhin 62 Stücke. Diese Verteilung ist mit einer Unschärfe behaftet, da sich in der HAB sicherlich unentdeckte Exemplare befinden. Übergreifendes Ziel der Auseinandersetzung mit dem Material ist das Kartieren dieser Katalogfülle und das Anlegen eines begehbaren Pfades durch sie. Dazu werden ca. 30 Kataloge eines repräsentativen Teilkorpus statistisch beschrieben, ihre Strukturdaten erhoben, d.h. virtuelle Inhaltsverzeichnisse erstellt, zudem erfolgt die sachsystematische Erschließung der Stücke. Neben diesen Arbeitsschritt tritt die mikroperspektivische Erforschung von einigen aussagekräftigen Einzelexemplaren und daran anknüpfend die Rekonstruktion von Besitzer- und Gelehrtenbiographien, flankiert von der elektronischen Edition komplementärer Quellen wie Briefkorrespondenzen. Ferner steht die Diskussion wissenschaftsgeschichtlich relevanter Fragen auf dem Programm und

schließlich die Darstellung und Recherchierbarkeit der selektierten Einzelkataloge in einer Datenbank.

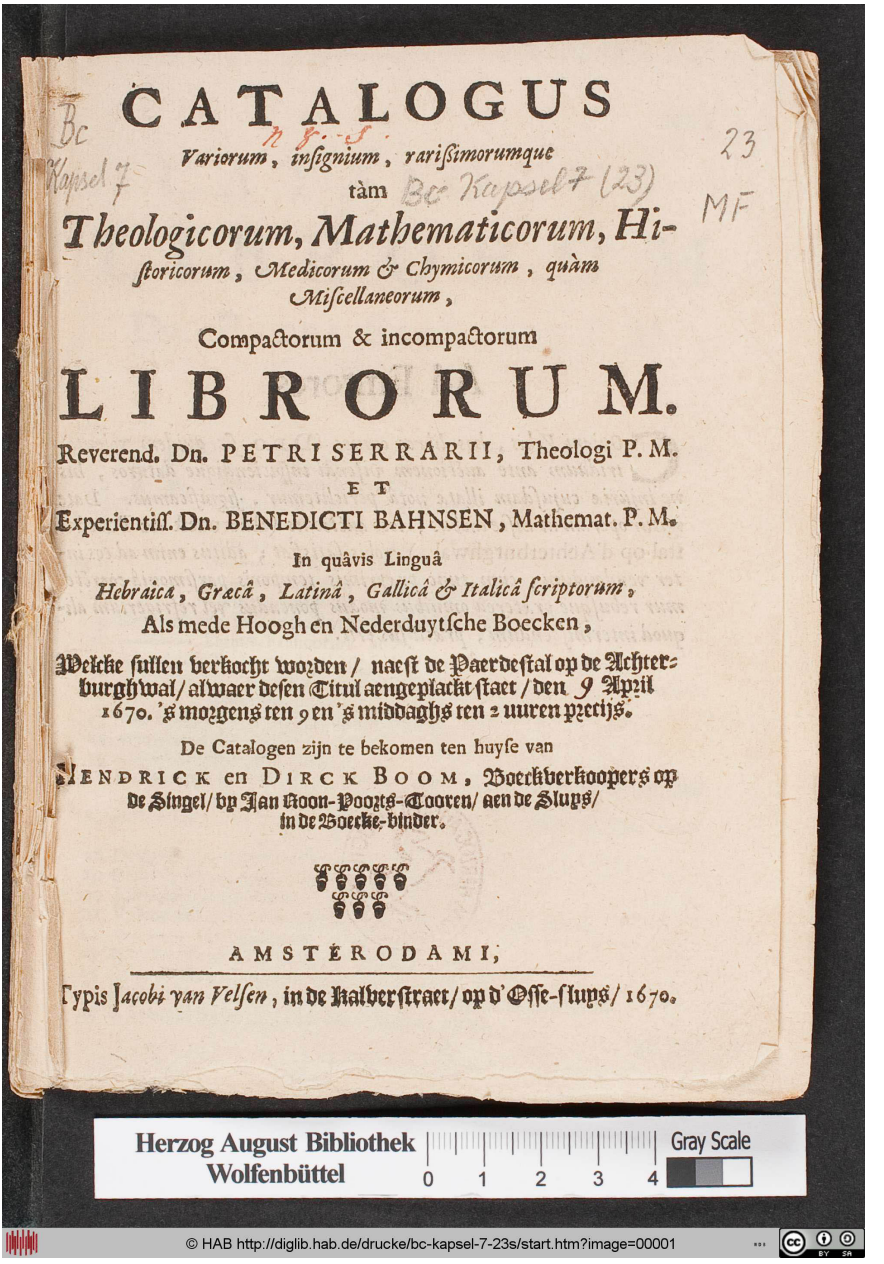
Der Katalog, um den es hier geht und der als Paradigma für die Erschließung frühneuzeitlicher privater Bibliotheken dient, listet den Buchbesitz des Chiliasten Benedikt Bahnsen (Hakelberg 2015b). Der aus Norddeutschland stammende Bahnsen lebte im 17. Jahrhundert und emigrierte in die Niederlande nach Amsterdam; wahrscheinlich musste er aus Glaubensgründen seine Heimat verlassen. Sein Sterbedatum ist das Jahr 1669. Die Annäherung an Bahnsen, der als Verleger, Buchhändler und Bücheragent, Mathematiker und Rechenmeister tätig war und als Autor hervorgetreten ist, gelingt über seine Bücherei. Diese ist physisch nicht mehr greifbar, dafür existiert sie als virtuelle Katalogaufstellung. Der gedruckte Auktionskatalog war bei den beiden Buchhändlern Dirk und Hendrik Boom in Amsterdam erhältlich; die Booms fungierten auch als Versteigerer. Zusammengebunden mit dem Bestandsverzeichnis der Gelehrtenbibliothek des Petrus Serrarius (1600-1669) kam er 1670 als Doppelkatalog auf den Markt. Laut Katalogdeckblatt (Abb. 1) wurden beide Büchersammlungen am 9. April 1670 beim Pferdestall auf dem Achterburgwall versteigert. Der Katalog liegt unbeschnitten in originaler Heftung vor. Die beiden Katalogteile, denen das gemeinsame Titelblatt vorgeheftet ist, sind zwar separat paginiert, ihre Lagen haben sich aber aufgrund einer irrtümlichen Heftung zum Teil ineinander verschoben.

Bahnsens Bibliothek umfasst insgesamt 2132 Lose, davon 2098 Nummern mit schätzungsweise 3000 Drucken und Handschriften, einschließlich mehrfach vorhandener Titel. Der Katalog verzeichnet zuerst gebundene, dann ungebundene Bücher. Die gebundenen Bücher sind – wie häufig in den zeitgenössischen Katalogen – in fünf Sachgruppen bzw. Abteilungen geordnet und innerhalb dieser nach Formaten unterteilt und nummeriert. Auf die Abteilung *Theologie* (I) folgen *Alchemie und Medizin* (II), *Mathematik und Geschichte* (III), *Verschiedenes* (IV) und *Handschriften* (V). Die ungebundenen Bücher sind nicht nummeriert, sondern nach Autoren bzw. Sachtiteln alphabetisch geordnet. Die restlichen 134 Lose sind in einer den Abschluss des Katalogs bildenden und mit ›Alderley Land-Caerten‹ überschriebenen Sonderabteilung gelistet. Sie verweisen auf 29 Land-, Himmelskarten, geometrische Grundrisse, eine Himmelskugel und einen Erdglobus, mathematische Instrumente, 70 Pakete unbestimmten Inhalts, weitere Varia sowie eine hohe Stückzahl ungebundener Exemplare von Werken, die Bahnsen verlegt hat und für den Vertrieb offenbar vorrätig hielt.

2 Bibliographische Erschließungsmethode

Als Erschließungsgrundlage dient neben dem Digitalisat des Auktionskataloges eine Excel-Tabelle.[1] In die Tabelle werden sämtliche Titel aus dem Katalog übertragen.

[1] Für die Arbeit mit ihr greifen wir auf eine Handreichung zurück, die ursprünglich von Dietrich Hakelberg im Rahmen des Projekts *Frühneuzeitliche Gelehrtenbibliotheken* konzipiert wurde und die Katrin Schmidt und Jörn Münkner im Prozess der Titelerfassung angepasst haben.



CATALOGUS

Variorum, insignium, rarissimorumque

tàm

Theologicorum, Mathematicorum, Historicorum, Medicorum & Chymicorum, quàm Miscellaneorum,

Compactorum & incompactorum

LIBRORUM.

Reverend. Dn. PETRI SERRARII, Theologi P. M.

ET

Experientiss. Dn. BENEDICTI BAHNSEN, Mathemat. P. M.

In quâvis Linguâ

Hebraica, Græcâ, Latinâ, Gallicâ & Italicâ scriptorum,

Als mede Hoogh en Nederduytsche Boecken,

Welcke sullen verkocht worden / naest de Paerdestal op de Achterburghwal / alwaer desen Titul aengeplackt staet / den 9 April 1670. 's morgens ten 9 en 's middaghs ten 2 uuren precijs.

De Catalogen zijn te bekomen ten huyse van

HENDRICK en DIRCK BOOM, Boeckverkoopers op de Singel / by Jan Roon-Poorts-Tooren / aen de Slups / in de Boecke-binder.

AMSTERODAMI,

Typis *Jacobi van Velsen*, in de Halverstraet / op d'Osse-slups / 1670.

Abbildung 1: Bahnsen/Serrarius-Katalog Titelblatt, Screenshot, Sign.: HAB Wolfenbüttel: Bc Kapsel 7 [23]

Diese vorlagegemäße Erschließung geschieht vorrangig auf Ebene der Ausgabe bzw. auf Werkebene, weil eine Identifizierung der physisch zumeist nicht vorhandenen Exemplare bislang nicht möglich ist. Der bibliographische Detailnachweis erfolgt über die Recherche in Online-Katalogen. Wenn zu einem späteren Zeitpunkt Exemplare aus dem Bestand Bahnsens identifiziert werden können, sollen sie digitalisiert und im Zuge der Überführung aller in der Excel-Tabelle vorhandenen Daten in eine Webpräsentation ebenfalls dorthin verlinkt, abrufbar und recherchierbar gemacht werden.

Um jeden Datensatz eindeutig zu identifizieren, d.h. um den Rückverweis sowohl auf das Digitalisat als auch auf die Druckvorlage zu gewährleisten, wird jeweils zunächst die Image-Nummer erfasst (Abb. 2). Dabei handelt es sich um die Nummer der digitalisierten Seite, die rechts unten auf jeder Seite des Digitalisats angegeben ist. Es folgt die Angabe der Seite, die der Seite im gedruckten Katalog entspricht, auf der der jeweilige Titel steht. Als drittes erfassen wir die Nummer, womit die laufende Nummer des Titels im Auktionskatalog gemeint ist.

Unter der Rubrik *Q = Qualität der Erfassung* verstehen wir die Treffgenauigkeit der Recherche, also das ermittelte bibliographische Level. Vier Auswahlmöglichkeiten stehen hier zur Verfügung. Der Buchstabe *e* steht für *Exemplar* und würde gewählt, wenn wir tatsächlich das Exemplar ausfindig machen könnten, das sich im Besitz von Benedikt Bahnsen befand. Das ist unwahrscheinlich und bislang nicht vorgekommen. Der Buchstabe *a* meint die *Ausgabe*, das heißt das Exemplar konnte zwar nicht gefunden, dafür aber die genaue Ausgabe identifiziert werden, zu der der Titel gehört. *w* verweist auf *Werk* und wird verwendet, wenn die Ausgabe nicht eindeutig identifizierbar ist und mehrere in Frage kommen. *o* wiederum steht für *ohne Nachweis* und zeigt an, dass ein Nachweis des Titels auch nach umfassender Recherche nicht gelungen ist. Generell gilt, dass immer der Nachweis berücksichtigt wird, der das beste bzw. vollständige Katalogisat bietet. Im Fall der als Beispiel dienenden Losnummer 9 auf der ersten Seite des Katalogs, dem Titel *Lutheri Kerckenpostilla. Witt. 1563*, handelt es sich eben um die Ausgabe der Lutherischen Postille, die wir nachweisen können.

Die Anzahl der möglichen *Autoren* und sonstigen *beteiligten Personen* pro Loseintrag haben wir auf vier beschränkt. Der Name des Autors bzw. der beteiligten Person wird gemäß Ansetzungsform der GND (Gemeinsame Normdatei) wiedergegeben, so dass Eintragungen unter verschiedenen Namenformen einer Person vermieden werden. Die GND-Namenvariante wird zudem in einer Hintergrundtabelle hinterlegt, so dass nur sie als normierte Form ausgewählt werden kann. Generell befinden sich hinter den meisten Rubriken in der Tabelle Hintergrundlisten, in denen die zu benutzenden Namen, Orte etc. hinterlegt sind, um eine größtmögliche Vereinheitlichung der Daten zu erreichen.

Die Titeldaten werden in dreifacher Form aufgenommen (Abb. 3). Zum ersten gibt es die vorlagegemäße Abschrift des *Titels*, so wie er im Auktionskatalog gedruckt ist, zum zweiten den *bibliographierten Titel*, wobei der komplette Hauptsachtitel des

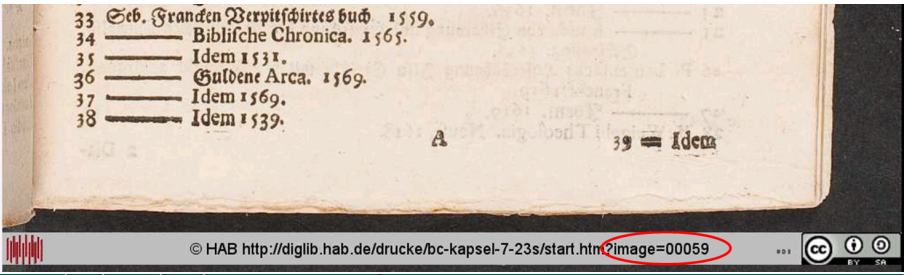


Image	Seite	Nr.	Q	Autor	Titel Vorlage
00059	1	8	a	Agricola, Conrad	Agricolae Deutsche Biblische Concordans. 1640.
00059	1	9	a	Luther, Martin	Lutheri Kerckenpostilla. Witt. 1563
00059	1	10	a	[Luther, Martin]	Idem [Lutheri Kerckenpostilla] Basel. 1528.

Abbildung 2: Ausschnitte aus Digitalisat und Excel-Tabelle, Screenshots

Titel Vorlage	Titel bibliographiert	Kurztitel
Agricolae Deutsche Biblische Concordans. 1640.	Concordantiae Bibliorum, Das ist Biblische Concordantz und Verzeich	Concordantiae Bibliorum
Lutheri Kerckenpostilla. Witt. 1563	KerckenPostilla || Dat ys/|| Vthleggínge der Euangelien an || den v#	Kerckenpostilla
Idem [Lutheri Kerckenpostilla] Basel. 1528.	Postille op die Epistelen ende Evangelien van allen sondaghen ende	Postille op die Epistelen ende Evangel
Idem [Lutheri Kerckenpostilla] Madgeburg. 1530. 1.2. tom.	Auslegung der || Episteln vnd Euangelien || vom Aduent an/ bis ||	Auslegung der Episteln und Evangelier
Idem [Lutheri Kerckenpostilla] Madgeburg. 1530. 1.2. tom.	Ausle=||gung der Euã=||gelien/ von Ostern || bis auffs Aduent/|| g	Auslegung der Evangelien von Ostern

Abbildung 3: Excel-Tabelle (Titelaufnahme), Screenshot, Ausschnitt

ermittelten Titels aus dem Online-Katalog übernommen wird, inklusive Versalien, Zeilenumbrüchen und typographischen Besonderheiten. Ist die Verfasserangabe mit dem Hauptsachtitel grammatikalisch verbunden oder enthält sie wichtige Zusatzinformationen, werden diese mit übernommen, was häufig bei Titeln aus dem VD17 und VD16 praktiziert wird. Als drittes wird, wenn möglich, ein *Kurztitel* angegeben, da insbesondere Titel aus dem VD16 lang und aufgrund vieler typographischer Besonderheiten unübersichtlich sein können.

Was den *Erscheinungsort* anbetrifft, so berücksichtigen wir maximal zwei Orte (Abb. 4). In der Hintergrundtabelle zum Ort werden auch die Koordinaten in Dezimalgrad eingetragen, die für die spätere Visualisierung mit einem Geo-Browser dienen. Des Weiteren werden maximal zwei *Drucker/Verleger* in der Tabelle vermerkt. Ihre Namen erfassen wir ebenfalls lt. GND-Ansetzungsform. Sind *Erscheinungsjahre* ermittelt worden, werden sie in eckigen Klammern angegeben und ggf. im Freitextfeld entsprechende Erläuterungen hinterlegt, wie auch möglichst bei allen recherchierten Daten.

Ort	Drucker/ Verleger1	Jahr
Frankfurt/M.	Hoffmann, Wolfgang	1640
Wittenberg	Krafft, Johann d.Ä.	1563
[Basel]	Hoochstraten, Johannes	1528

Format	Sachgruppe hist.	Sachbegriff	Gattungsbegriff
2°	Libri Theologici	Theologie	Konkordanz
2°	Libri Theologici	Theologie	Predigtsammlung
2°	Libri Theologici	Theologie	Predigtsammlung

Abbildung 4: Excel-Tabelle (Aufnahme: Ort, Drucker/Verleger, Jahr, Format), Ausschnitte, Screenshot

Die Erfassung des *Formats* erfolgt entsprechend der Angabe im gedruckten Katalog. Dort sind, wie erwähnt, die Titel nach Sachgruppen und innerhalb der Sachgruppen nach Formaten unterteilt. Das bibliographische Format wird in normierter Kurzform eingetragen. Das Format ist übrigens oft ein wertvoller Hinweis darauf, ob eine bestimmte Ausgabe eines Titels als möglicher Treffer in Frage kommt.

Hinter der Rubrik *Sachgruppe historisch* verbirgt sich die vorgegebene Sachgruppen- bzw. Abteilungsordnung im Katalog. So handelt es sich bei besagter *Lutherischer Postille* um einen Titel, der zu den *libri theologici* gehört. Die *Sach- und Gattungsbegriffe* basieren auf der Bestimmung, die sich anhand des jeweiligen Katalogisats des gesuchten Titels ergibt; sie entsprechen größtenteils den normierten Sach- und Gattungsbegriffen der AAD.[2] Ggf. muss ein Datensatz präzisiert und ergänzt werden, wenn im Katalogisat eines betreffenden Titels kein Gattungsbegriff vergeben wurde oder es gar keinen geeigneten Begriff gibt. Im Fall der Lutherischen Postille wird dieses Manko ersichtlich: Im Normset der AAD gibt es keinen generellen Sachbegriff *Theologie*, diesen haben wir für die Postille ergänzt, die Leerstelle also selbständig ausgeglichen. Eine Abweichung von den AAD-Vorgaben sollte allerdings die Ausnahme bleiben.

Was die Rubrik *Medium* anbetrifft, wurden bislang die Angaben *Druck*, *Handschrift* und *Sache* verwendet. *Sache* indiziert Sammlungsobjekte wie Globen, Instrumente oder ähnliches. Es werden maximal zwei *Sprachen* aufgenommen, und zwar jeweils im Buchstaben-Sprachcode nach ISO 639-2. Unter *Form* ist zwischen gebunden und ungebunden zu wählen.

Der *Nachweis* eines Loses und damit eines Titels oder einer Ausgabe setzt sich zum einen durch den in Kurzform angegebenen Namen des Online-Kataloges wie z.B. VD17, VD16, GBV etc. und der dazu gehörenden ID zusammen. Das ist der Fall, wenn eine Ausgabe eindeutig identifizierbar ist. Kommen mehrere Ausgaben in Betracht, wird »Ausgabe nicht bestimmbar« gewählt und es werden in einem Freitextfeld die

[2] Arbeitsgemeinschaft Alte Drucke beim GBV.

IDs von bis zu drei in Frage kommenden Ausgaben angegeben. Kommen mehr als drei Ausgaben in Betracht, hinterlegen wir momentan im Freitextfeld einen Suchbefehl für den jeweiligen Online-Katalog, so dass alle möglichen Ausgaben angezeigt werden. Zudem können in der Freitextspalte Bemerkungen aller Art geparkt werden, z.B. welche Funktion eine beteiligte Person am Werk hatte oder Erläuterungen, wo bestimmte Informationen ermittelt wurden.

Unter der Rubrik *Digital* werden schließlich die URL oder URN des Digitalisats einer identifizierten Ausgabe hinterlegt. Kommt bei der Recherche mehr als eine Ausgabe in Betracht, wird hier der Link zu einem repräsentativen Digitalisat verankert. In der Freitextspalte wird entsprechend vermerkt, zu welcher möglichen Ausgabe das Digitalisat gehört. Falls kein Digitalisat nachgewiesen ist, im VD17 aber Schlüsselseiten vorhanden sind, werden auch diese verlinkt, da sie eine gute Informationsquelle darstellen. Handelt es sich um einen Link zu einer Schlüsselseite, bekommt die URL ein Präfix, damit bei der Datenbank-Darstellung der jeweilige Link mit *Digitalisat* oder *Schlüsselseite* eingeleitet wird. In der Rubrik *Onlinebiographien* können die URLs beispielsweise zu Einträgen in der ADB (Allgemeine Deutsche Biographie) oder ggf. auch zu Wikipedia hinterlegt werden, wo weitergehende Informationen zum Verfasser des jeweiligen Titels abrufbar sind.

3 Entwicklung einer digitalen Infrastruktur

Für die digitale Präsentation rekonstruierter Bibliotheken gibt es keine fixe und etablierte Lösung. Die bibliothekarische Infrastruktur ist nur bedingt geeignet, weil sie für vorhandene Bestände und autoptische Erschließung ausgelegt ist. Fehlt eine physisch-konkrete Vorlage, wie in diesem Projekt, so kann man zwar einen virtuellen Exemplarsatz im Katalog verzeichnen, der den Benutzenden anzeigt, dass ein bestimmtes Buch in einer bestimmten Bibliothek vorhanden war, viele Informationen des Altkatalogs gehen aber verloren. Auch Fälle ungewisser Zuordnung, die im Bahnsen-Katalog in Fülle vorkommen, sind nur schwer abzufangen. Eine andere Infrastruktur, in die sich das hier vorgestellte Vorhaben theoretisch integrieren ließe, bietet diejenige für digitale Editionen. Ein Präzedenzfall ist das an der HAB edierte Bücherinventar der Elisabeth von Calenberg (1510–1558) (Bücherinventar Calenberg). Ein solches Vorgehen bleibt zwangsläufig eng an die Vorlage gebunden. Auswertungsfunktionen, die eine Annäherung und Durchdringung der Sammlung in ihren verschiedenen Facetten erlauben – räumlich, zeitlich, prosopographisch, inhaltlich –, sind nur schwer zu integrieren. Die digitale Rekonstruktion nicht erhaltener Bibliotheken erfolgt daher meist mit einer eigenständigen Datenbank, die über ein ad hoc geschriebenes Webinterface ausgewertet wird. Die Präsentation ist damit gegenstandsspezifisch und nicht nachnutzbar. Im Rahmen eines Unternehmens, das nicht eine einzelne

bedeutende, sondern mehrere exemplarische Bibliotheken rekonstruiert, stellt sich die Frage nach einer generischen, nachnutzbaren Lösung umso mehr. Die erhobenen Daten sind hierfür gut geeignet, weil sie qualitativ hochwertig, stark strukturiert und mit Normdatenverknüpfungen für Personen und Orte angereichert sind.

Eine Software für die Präsentation bibliographischer Daten von einer einstmaligen Bibliothek als Website ist eine Anforderung, die im Kontext von bibliotheksbezogenen Forschungsprojekten immer wieder begegnet, wenn man vom Spezialfall des Auktionskatalogs einmal abstrahiert. Alle Arten von Bücherverzeichnissen und Katalogen können zur Grundlage einer solchen Präsentation werden: zu denken ist an Nachlassinventare, Neuerscheinungslisten sowie thematische oder personenbezogene Bibliographien. Sehr ähnlich sind die Anforderungen, wenn Bibliotheken aus Provenienzdaten rekonstruiert werden sollen, wobei nicht ein Katalog, sondern die erhaltene Sammlung den Ausgangspunkt bildet; anstelle von Verweisen auf einen Altkatalog können dann Links in den jeweiligen OPAC angezeigt werden. Schließlich sind auch andere Szenarien denkbar, in denen weder ein Altkatalog noch Provenienzdaten vorliegen. So führt auch die Auswertung von Zitationen oder von anderen Lektürezeugnissen zur Rekonstruktion ehemaliger Sammlungen. Auch bei überlieferten und katalogmäßig erschlossenen Sammlungen ist die Darstellung als Website zur inhaltlichen Auswertung manchmal sinnvoll, etwa wenn ein Teilbestand mit eigener Geschichte erforscht werden soll.

Ausgangspunkt für die Gestaltung der Oberfläche für die Bahnsen-Webpräsentation war die Rekonstruktion der Privatbibliothek des Mathematikers, Astronomen und Arztes Duncan Liddel (1561–1613) im Rahmen einer Kooperation zwischen der Universität Aberdeen und der Herzog August Bibliothek. Die Bücher, die Liddel in seiner langen akademischen Tätigkeit auf dem Kontinent sammelte und dem Marischal College in Aberdeen hinterließ, konnten anhand von Provenienzforschungen identifiziert werden. Es entstand eine Website, die die Sammlung in mehreren strukturierten Listen präsentiert; die einzelnen Datensätze sind mit Zusatzinformationen angereichert (Privatbibliothek Liddel). Die Herausforderung beim gegenwärtigen Projekt war es, ein Programm zu entwickeln, das Webseiten mit strukturierten Listen unabhängig vom jeweiligen Datencorpus erstellen kann. Die Eingabe sollte dabei über ein intuitiv bedienbares Format erfolgen und die Darstellung derart konfigurierbar sein, dass alle erfassten Kategorien auf Wunsch zur Grundlage einer Liste werden können. Im Ergebnis entstand ein Set aus PHP-Skripten, das zusammen mit den notwendigen Dokumentationen unter einer freien Lizenz im Netz veröffentlicht ist.[3] Es handelt sich um ein browserbasiertes Programm, das auf einem lokalen oder einem öffentlichen Server genutzt werden kann. Als Output liefert es HTML-Seiten, die im Browser betrachtet und als Datei heruntergeladen werden können. Das Webdesign ist

[3] <https://github.com/hbeyer/liddel-tool>.

zurückhaltend unter Verwendung des CSS-Frameworks Bootstrap gestaltet. Eine nachträgliche Anpassung an die Formatierungen des eigenen Webauftritts ist somit gut möglich. Zentrales Element ist eine Menüleiste, die für jede berücksichtigte Kategorie ein aufklappbares Inhaltsverzeichnis enthält (Abb. 5). Weil die Menüleiste auch beim Scrollen sichtbar bleibt, ist der Zugriff auf jede Stelle von überall her gewährleistet. Die Inhaltsverzeichnisse enthalten in runden Klammern die Anzahl der Datensätze zu jedem Abschnitt.

Im Kopfbereich der Sammlung werden Metainformationen zur Sammlung oder zum zugrunde liegenden Katalog ausgegeben. Der Link *Anzeige in Vorlageform* bewirkt, dass statt der bibliographierten Daten für jeden Eintrag eine Transkription des Eintrags im Altkatalog angezeigt wird. Zu jedem Datensatz können Zusatzinformationen eingeblendet werden. Diese werden je nach Verfügbarkeit entsprechender Daten generiert:

- Link auf die Seite im Digitalisat des Altkatalogs
- Link auf einen Datensatz für die Ausgabe in einem bibliographischen Nachweissystem (VD16–18, Verbundkataloge u. a.)
- Link auf das Originalexemplar im OPAC der besitzenden Institution
- Link auf einen Normdatensatz für das enthaltene Werk
- Link auf ein Digitalisat
- Freitextfeld mit Zusatzinformationen

Möglich wird so eine Darstellung, die unabhängig von der Datenquelle ist: Provenienzdaten können ebenso verarbeitet werden wie obskure oder nicht einer Ausgabe zuzuordnende Einträge in Altkatalogen. Zu jeder erfassten Kategorie kann eine eigene Liste generiert werden. Das Programm bildet automatisch Kategorien, wobei leere Felder zu einem Eintrag unter *ohne Kategorie* führen. Dabei gilt insgesamt das Prinzip, dass Gleiches gleich benannt werden muss, um in einem Eintrag zusammengeführt zu werden.

Die starke Normierung der Daten und ihre Anreicherung mit Normdaten ermöglicht es, Visualisierungen und Zusatzinformationen einzubauen.

Zwei auf JavaScript basierende Features sind Wortwolken und Kreisdiagramme (Abb. 6).

Die Datengrundlage ist frei wählbar; in der Regel eignen sich nur bestimmte Kategorien für jede Darstellungsform. Wortwolken sind für Felder mit vielfältigem, aber nicht individuellem Inhalt wie Person, Ort, Drucker, Schlagwort, Gattung geeignet. Kreisdiagramme dienen der Relationierung von wenigen, häufig vorkommenden Werten wie Sprache, Rubrik, Medientyp und Gattung (Abb. 7).

Die Darstellung von Ortsdaten erfolgt extern in dem von *DARIAH* als offenes Angebot gehosteten *Geo-Browser* (Abb. 8). Das Programm erzeugt hierfür eine CSV-Datei mit den Geoinformationen, die in das Repository geladen und dort referenziert

Abbildung 5: Menüleiste des browserbasierten Datenvisualisierungsprogramms mit Inhaltsverzeichnis/ Datensätze, Screenshot

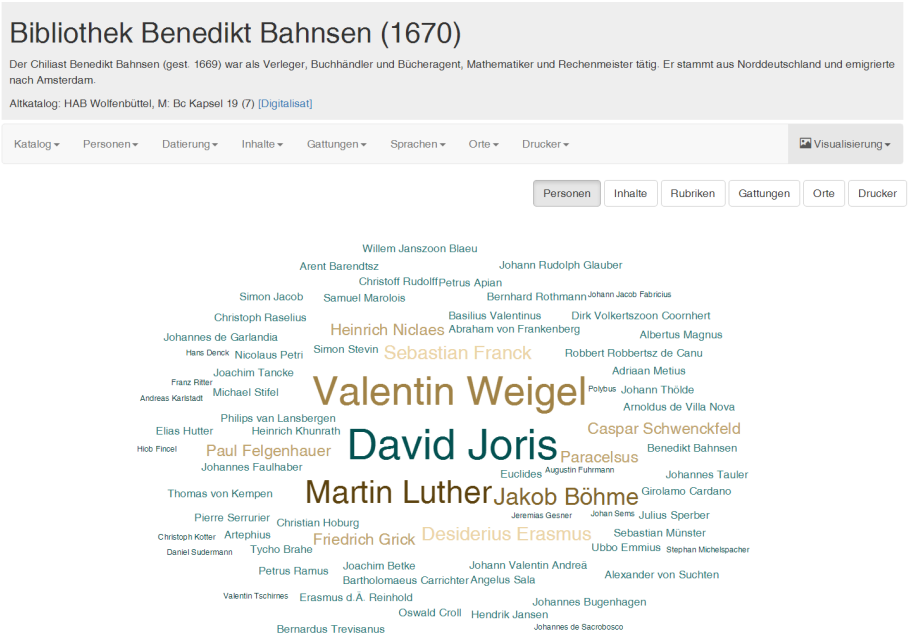


Abbildung 6: Wortwolken, erstellt mit Hilfe des Datenvisualisierungsprogramms, Screenshot

werden kann. Der *Geo-Browser* bietet nicht nur eine räumliche Darstellung der Entstehungsorte von Sammlungsstücken, sondern auch eine Zeitleiste mit graphischer Darstellung der quantitativen Verteilung.

Eine Anreicherung der Personendaten erfolgt mit Hilfe der GND-Nummer. Auf die Erfassung weiterer biographischer Daten wird verzichtet, stattdessen werden so genannte BEACON-Dateien auf ein Vorkommen der GND-Nummer durchsucht, was die Generierung von Links zu personenspezifischen Informationen ermöglicht. Das im Rahmen der deutschsprachigen Wikipedia entwickelte Dateiformat zur Meldung von Personeninformationen wird von zahlreichen Anbietern unterstützt, darunter biographische Standardwerke, Professorenkataloge, Bilddatenbanken und Briefeditionen. Die Einbindung erfolgt über ein Informationssymbol neben jedem Namen auf der Seite *Personen*.

Ein großer Vorteil dieser Präsentationsweise ist, dass die Bezugnahme auf bibliographische Nachweissysteme explizit gemacht wird. Da die Projektakteure kein Exemplar physisch vorliegen haben, bei dem sie die bibliographischen Angaben nachprüfen können, zitieren sie einzelne Datensätze aus Nachweissystemen, die persistent in

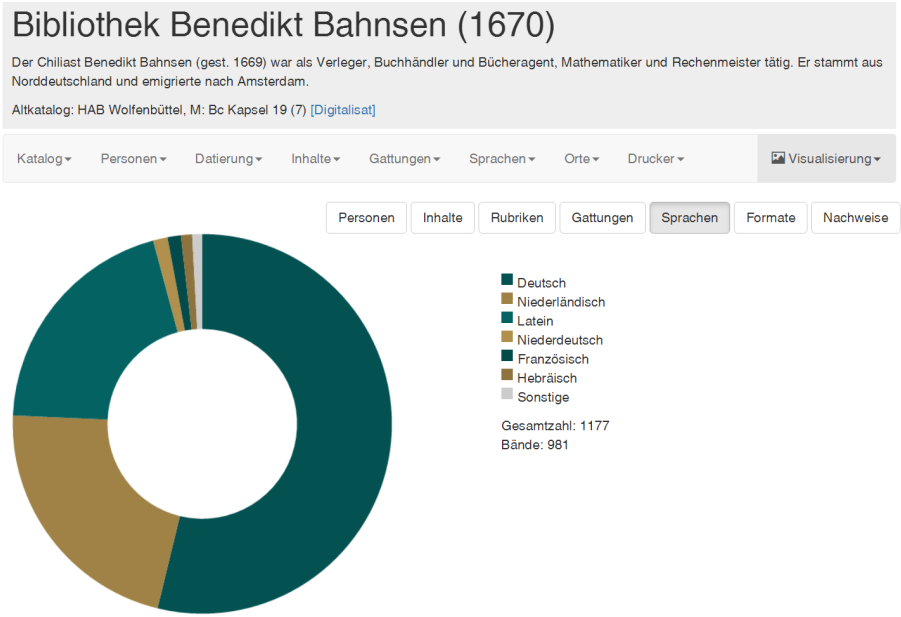


Abbildung 7: Kreisdiagramm, erstellt mit Hilfe des Visualisierungsprogramms, Screenshot

der vom Programm generierten Präsentation verlinkt sind. Es ist daher nicht notwendig, alle Details der bibliographischen Aufnahme bei der Rekonstruktion einer Bibliothek aufzunehmen, da diese verlässlicher und auf Basis von Regelwerken in den Redaktionen der Nachweissysteme erfasst werden. Die Zuordnung zwischen dem Eintrag im Altkatalog und der Ausgabe, der das Exemplar angehörte, ist die eigentliche Erschließungsleistung. Ihre Zuverlässigkeit hängt maßgeblich mit der Interpretierbarkeit des Katalogeintrags zusammen. Zweifelsfälle entstehen insbesondere bei sehr verbreiteten Werken mit mehreren Auflagen in einem Jahr oder solchen, die wegen ihrer Bekanntheit ungenau bezeichnet werden, z. B. als »Bibel deutsch«. Neben der Möglichkeit, das Feld für die Ausgabe frei zu lassen, kann auch ein Link zu einem Normdatensatz für ein Werk eingefügt werden, ein mögliches Nachweissystem hierfür ist die Gemeinsame Normdatei im Katalog der Deutschen Nationalbibliothek. Hinweise auf mehrere mögliche Ausgaben können im Freitextfeld vermerkt werden. Folgen diese einer festgelegten Syntax (Sigle, Leerzeichen, Identifier), so generiert das Programm automatisch eine Verlinkung der einzelnen Datensätze.

Das vorgestellte Programm trennt strikt zwischen dem Import und der Weiterverarbeitung von Daten. In einem ersten Schritt werden die zu transformierenden Daten

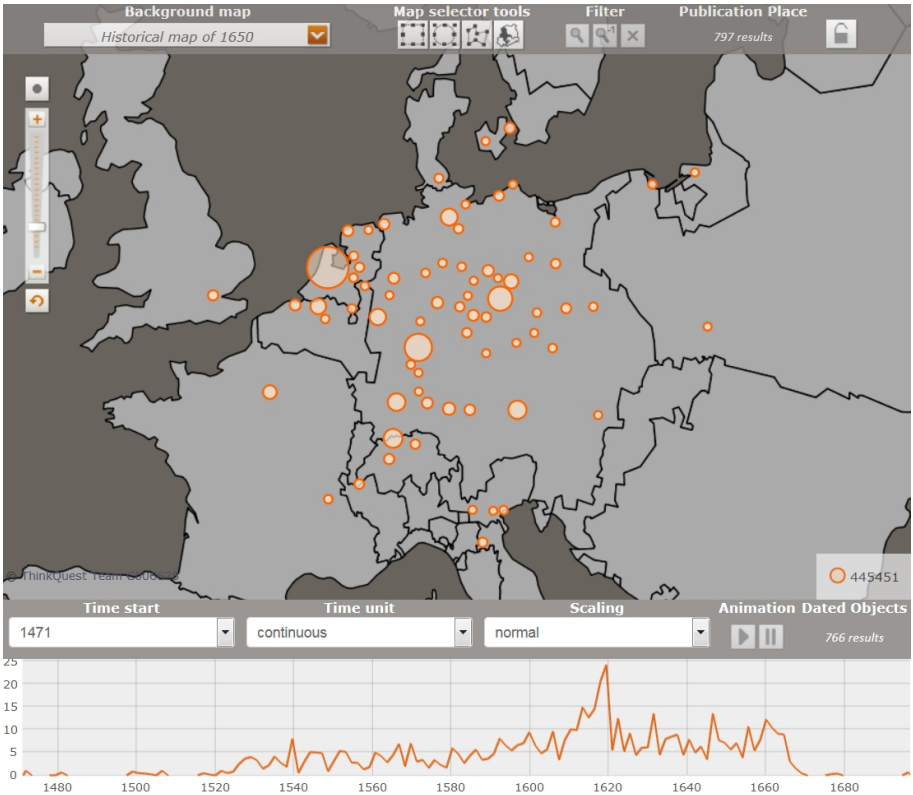


Abbildung 8: Druckorte und Erscheinungsjahre der Bahnsen-Bibliothek, visualisiert mit dem Geo-Browser von DARIAH-DE

in ein internes Modell umgerechnet und in dieser Form gespeichert. Hierdurch wird es sehr einfach, zusätzliche Eingabeformate zu implementieren. Im Projekt selbst konnten die in Excel vorliegenden Daten über den Umweg einer Datenbank geladen werden. Weil dies die Nachnutzung erschwert, wurde inzwischen eine Möglichkeit geschaffen, Dokumente im CSV-Format[4] zu verarbeiten. Hierdurch wird es möglich, Datensammlungen mit einem Tabellenkalkulationsprogramm zu erstellen und direkt zu transformieren. Voraussetzung war, dass sämtliche Informationen, die zu einem Datensatz gehören, auf einer Tabellenzeile untergebracht werden mussten. Für

[4] Comma-separated values: ein einfaches Tabellenformat, das sowohl mit einem Programm wie Excel als auch mit einem Texteditor bearbeitet werden kann.

mehrere Werte gibt es daher eine begrenzte Anzahl von Feldern (Personen, Orte) oder die Möglichkeit, mehrere Angaben durch Semikolon getrennt in einem Feld unterzubringen (Schlagwörter, Gattungen, Sprachen). Nutzt man das CSV-Format zur Eingabe, muss man sich daher auf je bis zu vier Autoren und Beiträger sowie zwei Orte beschränken, obwohl vom Programm mehr dargestellt werden könnten. Das Problem entfällt, nutzt man den ebenfalls möglichen XML-Upload. Hierfür wird ein XML-Schema bereitgestellt, das die Erschließung vereinfacht und lenkt. So ist eine Liste der zu verwendenden Siglen für bibliographische Nachweissysteme hinterlegt, die in gängigen XML-Editoren ein Auswahlmenü erzeugt.

Weil die manuelle Metadatenextraktion aus den Nachweissystemen langwierig ist, erscheint eine Anbindung an bibliographische Schnittstellen sinnvoll. So könnten bibliographische Daten (zumindest Titel, Personen, Ort, Drucker, Erscheinungsjahr) theoretisch auch vom Programm nachgeladen werden, wenn Nachweissystem und Identifier bekannt sind. Weil das angesichts der Vielzahl von Nachweissystemen, die nicht alle über offene XML-Schnittstellen verfügen, nur mit erheblichem Aufwand zu leisten wäre, und weil die Normierung, etwa von Ortsnamen, auf diese Weise nicht gewährleistet ist, wurde das nicht versucht. Implementiert wurde aber eine Möglichkeit zur Datenübernahme aus Literaturverwaltungsprogrammen wie *Citavi* oder *Zotero*. Exportiert man die Datensätze aus dem Literaturverwaltungsprogramm in das XML-basierte MODS-Format, werden sie vom Programm bei der Eingabe in das eigene XML- oder CSV-Format umgewandelt. Eine Nachbearbeitung ist sowohl im Literaturverwaltungsprogramm als auch in der XML- oder CSV-Datei möglich. Diese vereinfachte Datenerfassung ist vor allem dann interessant, wenn es nicht darum geht, die Anordnung der Sammlung in einem Altkatalog wiederzugeben, sondern wenn Provenienzdaten oder Sammlungsdaten aus anderen Quellen verarbeitet werden sollen.

Dieselbe Flexibilität wie beim Import weist das Programm bezüglich des Datenexports auf. Die Datenhaltung als PHP-Objekte vereinfacht die Programmierung von Transformationsroutinen für verschiedenste Formate. Zunächst kann das Programm die Daten in den möglichen Uploadformaten CSV und XML wieder ausgeben und ermöglicht so die Weiterverarbeitung oder das Nachbearbeiten für eine erneute Transformation. Weil diese Formate proprietär und daher nicht für die Langzeitarchivierung geeignet sind, empfiehlt sich die Transformation in einen etablierten Web-Standard. Hierbei sollen zwei Wege beschritten werden: Die Auszeichnungssprache TEI-P5 der *Text Encoding Initiative (TEI)* bildet den anerkanntesten und verbreitetesten Standard für digitales Edieren im engeren Sinne.[5] Wegen der großen Verbreitung und

[5] Neben der Wiedergabe von Vorlagen erlaubt sie auch die Anreicherung der Texte mit Metadaten sowie deren separate Verwaltung etwa in der Form von Personen- oder Werklisten. Sämtliche im Projekt erhobenen Daten können daher auch dann in einem TEI-Dokument untergebracht werden, wenn sie nicht einem Altkatalog entnommen sind.

minutiösen Dokumentation eignet sich das TEI-Format gut, um die langfristige Interpretierbarkeit der Daten sicher zu stellen. Die extrem breite Einsetzbarkeit der TEI bringt aber auch das Problem mit sich, dass keine standardisierten Anwendungen zur Anzeige aller möglichen TEI-Dokumente existieren. Das Projekt arbeitet daher an einem projektspezifischen XML-Schema auf Basis der TEI, zum dem ein Transformationsszenario zur Anzeige der Datensammlung als digitale Edition erstellt werden kann.

Ein anderer Weg, die Daten unabhängig von der bestehenden Anwendung nutzbar zu halten, liegt in der Transformation in semantische Daten nach den Regeln des *Resource Description Framework* (RDF). Hierbei werden die Daten auf der untersten möglichen Ebene in zahlreiche maschinenlesbare Tripel nach dem Muster ›Subjekt – Prädikat – Objekt‹ zerlegt. Jeder dieser Tripel ist anschließend unabhängig von seinem Kontext interpretierbar und maschinell auswertbar. Zunächst muss dabei zu jeder Ressource die zu beschreiben ist (Subjekt), eine URL gebildet werden. Dieser URL kann eine weitere URL oder auch eine Zeichenkette für das Objekt zugewiesen werden. Die Art der Beziehung wird durch das Prädikat ausgedrückt, das ebenfalls in Form einer URL codiert ist. Um die Daten zu historischen Bibliotheken in dieser Form ausdrücken zu können, ist es zunächst notwendig, die zu beschreibende Ressource genauer zu definieren (je nach Betrachtungsweise ist es das Sammlungsobjekt selbst, der Katalogeintrag oder das Besitzverhältnis) und eine projektinterne Konvention zur Zuweisung von URLs festzulegen. Für die Beschreibung dieser Ressource ist vor allem ein Set aus Prädikaten notwendig. Diese können und sollten möglichst aus bekannten und im Web verwendeten Ontologien entnommen werden, um die Nutzung der Daten außerhalb des Projektkontextes zu erleichtern. Hat man auf diese Weise für jedes Feld der Ausgangsdatensätze eine Beschreibungsform in Tripeln etabliert, lässt sich die automatische Umwandlung der Daten zu einer historischen Bibliothek in RDF einfach in das Programm integrieren. Der Nutzen einer RDF-Transformation besteht zum einen in der langfristigen Sicherung, zum anderen in der Kombinierbarkeit mit anderen semantischen Daten im Web. Gerade weil das gegenwärtige Projekt nicht auf die Beschreibung von überlieferten Medien, sondern auf die Beziehungen von Sammlern und ihren nicht erhaltenen Sammlungsstücken ausgerichtet ist, eignet sich ein solches relationsorientiertes Vorgehen, um zukünftige Forschung an den Daten zu ermöglichen.

Die Entwicklung einer Suchfunktion wurde im Projekt erst im zweiten Schritt angegangen. Durch die Verfügbarkeit aller Titel als Liste ist eine rudimentäre Volltextsuche schon mit der im Browser eingebauten Funktionalität möglich; die strukturierte Anzeige nach mehreren Facetten ersetzt die Suche in einzelnen Feldern zudem sehr effektiv. Was den einfachen Rahmen sprengt, ist aber zum einen die Suche über mehrere rekonstruierte Bibliotheken hinweg, zum anderen die Kombination mehrerer Suchkriterien. Für diese Anforderungen erwies sich die nachträgliche Übernahme der

Daten in eine auf dem Markt verfügbare Suchmaschinen-Software als effizienteste Lösung. Gewählt wurde hierfür *Apache Solr*, ein Open-Source-Framework um den Suchindex Lucene, das an der HAB bereits an anderen Stellen im Einsatz ist. Hierfür werden die Daten jeder einzelnen rekonstruierten Bibliothek in ein von Solr direkt lesbares XML-Format umgerechnet, auch hierfür müssen die Daten in eine ›flache‹ Form überführt und hierarchische Verknüpfungen aufgelöst werden. Das Solr-Framework stellt eine Fülle von Funktionalitäten zur Verfügung, wie sie aus modernen Discovery-Services bekannt sind. Dazu gehören neben vielem anderen eine leistungsfähige Volltextsuche über alle Felder, eine feldspezifische Suche in beliebiger Kombination, das Definieren von Filtern (besonders nützlich, will man auf einzelne Bibliotheken einschränken), eine Bereichssuche für Jahreszahlen, Rechts- und Linkstrunkierungen und die Möglichkeit der unscharfen Suche, die besonders für die ungeregelte Orthographie früherer Jahrhunderte von großem Wert ist. Besonders interessant ist die Möglichkeit der Facettierung, hierbei können für die Treffermenge (bzw. den Gesamtdatenbestand) die vorkommenden Werte der einzelnen Felder sichtbar und aufrufbar gemacht werden. Auf diese Weise entsteht eine Suchmaschine für die zugrunde liegenden Bibliotheksrekonstruktionen, die unmittelbar zur Datenanalyse genutzt werden kann.

Die Integration der Daten in eine solche Suchmaschine ist prinzipiell geeignet, die vom Programm generierte Ansicht als statische Website zu ersetzen. Das gilt besonders dann, wenn man es mit größeren Datenmengen als einer kleineren vierstelligen Zahl an Titeln zu tun hat. Weil von großen Trefferlisten immer nur ein Teil angezeigt und die Suche selbst sehr performant ausgeführt wird, sind dem Mengenwachstum kaum Grenzen gesetzt. Das Programm, das im Rahmen des Projekts veröffentlicht wurde, wird dadurch aber nicht obsolet. Wichtig bleibt es als Tool für die Normierung, Anreicherung und Transformation von Daten, die mit Solr lediglich dargestellt werden. Zudem kann die Solr-Instanz nicht für die externe Nachnutzung bereitgestellt werden. Für externe Nutzer ist daher das Visualisierungstool eine einfache Möglichkeit zur Generierung von Webseiten zu historischen Büchersammlungen nach Kriterien, die für eine Vielzahl von Projekten relevant sind.

4 Kombination von ›traditioneller‹ Altbestandserschließung und Digital Humanities

Der *Forschungsverbund Marbach Weimar Wolfenbüttel* (MWW), bestehend aus dem Deutschen Literaturarchiv Marbach, der Klassikstiftung Weimar und der Herzog August Bibliothek Wolfenbüttel, vereint ›traditionell‹ geisteswissenschaftliche wie Digital Humanities-Forschungsprojekte unter seinem Dach. Dazu gehört das diesem Beitrag zugrunde liegende Projekt zur Erschließung frühneuzeitlicher Gelehrtenbiblio-

theken vermittels Auktionskatalogen aus dem Bestand der Herzog August Bibliothek. Der ursprüngliche Zuschnitt des Projekts sah die Erschließung der Auktionskataloge ohne Unterstützung der Digital Humanities (im Folgenden wird die Abk. ›DH‹ verwendet) vor, da der Schwerpunkt der DH-Projekte im Forschungsverbund auf dem Aufbau einer gemeinsamen digitalen Infrastruktur für die wissenschaftliche Nutzung der (digitalen) Bestände der Verbundeinrichtungen liegt; deswegen stehen nur begrenzte zeitliche und personelle Ressourcen für die Zusammenarbeit mit den einzelnen Forschungsprojekten zur Verfügung. Trotz dieser Ausgangslage wurden früh Synergieeffekte ersichtlich, wie sie aus der Zusammenarbeit zwischen ›traditionellen‹ und ›digitalen‹ Projektformen resultieren. Einerseits konnten die DH-Projekte das Anforderungsprofil der entstehenden virtuellen Forschungsumgebung durch den Projekt-Usecase spezifizieren. Anderseits zeichnete sich ab, dass die Erschließungsarbeit durch die Anwendung von DH-Methoden effizienter gestaltet, ein auf die erhobenen Erschließungsdaten zugeschnittenes Präsentationsangebot aufgebaut und Verfahren für die Datenanalyse gemeinsam evaluiert und angewendet werden konnten (vgl. die Ausführungen unter Punkt 3).

Die Zusammenarbeit zwischen Digital Humanists und Forschern aus anderen geisteswissenschaftlichen Disziplinen birgt nach wie vor ein gewisses Konfliktpotential, denn nicht jeder sieht in der Nutzung computergestützter Verfahren einen Mehrwert. Sicherlich gehört auch ein gewisses Maß an Vertrauen dazu, sich von gewohnten Arbeitsweisen zu entfernen und neue Wege zu beschreiten, bei denen man auf einen technischen Begleiter angewiesen ist. Auf Seiten der DH besteht fraglos ein Spannungsverhältnis zwischen der gewünschten Rolle als gleichberechtigter Partner in Forschungsprojekten und der häufig zugewiesenen Bedeutung als reiner Serviceentwickler und Dienstleister. Wahrscheinlich wird sich dieses Dilemma mit der zunehmenden Etablierung und Professionalisierung der DH mittelfristig auflösen, Bedingung dafür sind aber die Bereitschaft, sich auf die neuen Geist-Technologie-Partnerschaften einzulassen, entsprechende Angebote und die Konsolidierung von entsprechend nachhaltigen Strukturen in den Universitäten und Gedächtniseinrichtungen (Sahle 2015, passim; Kaden 2013).

Für die Erschließung frühneuzeitlicher Auktionskataloge lag es nah, traditionelle Verfahren der Altbestandserschließung mit Methoden der DH zu kombinieren. Im Folgenden seien ausgewählte Aspekte der Gestaltung und Ausformung dieser Zusammenarbeit und ihre Ergebnisse vorgestellt. Des Weiteren geht es um eine Erörterung, inwieweit die Methoden der DH für die Erschließung der frühneuzeitlichen Auktionskataloge einen Mehrwert leisten. Folgende Annahmen bilden den Ausgangspunkt:

- Erschließungsprojekte beinhalten heutzutage immer digitale Komponente(n), seien es Digitalisate, die Integration der Erschließungsdaten in neue oder bestehende Nachweissysteme und/oder Projektdatenbanken.

- Gedächtniseinrichtungen können mittlerweile bei Erschließungsprojekten sowohl eine digitale Infrastruktur (die eigene oder eine externe) als auch etablierte Workflows einbringen. Je größer das digitale Angebot, desto mehr Möglichkeiten bieten sich für Erschließungsprojekte, ihre Daten zu normalisieren und automatisiert mit Wissen anzureichern (Neuroth et al. 2009, 161-169).
- Die digitale Erschließung und Präsentation des Bestandes ist eine dauerhafte Aufgabe von Gedächtniseinrichtungen. Ergänzend kommt es durch den Aufbau von digitalen Angeboten und die Erwerbung digitaler Medien zum Aufbau von digitalen Beständen, die keine physische Entsprechung mehr in den analogen Beständen besitzen.
- Die DH stellen sowohl Tools als auch Methoden zur Verfügung, die die Altbestandserschließung in mehreren Bereichen unterstützen.
- Die Anwendung von Methoden der DH darf auf keinen Fall unreflektiert, mit blindem Vertrauen auf die Möglichkeiten von technischen Verfahren erfolgen, sondern muss mit den Zielen des Projektes und den zur Verfügung stehenden bzw. zu benutzenden digitalen Materialien eng abgestimmt werden.

Das Erschließungsvorhaben von frühneuzeitlichen Auktionskatalogen, das eine Vielzahl von Exemplaren umfasst, ist aufgrund der strukturellen Heterogenität, der geringen Standardisierung und der Informationsmenge keine kleine Herausforderung. Auktionskataloge weisen eine große Bandbreite an Inhaltsstrukturierungen auf, oft gepaart mit vielfältigen Benutzungsmerkmalen, was vielfach im unikalen Charakter der Kataloge resultiert (Vogeler 2015, passim). Dies stellt den Bearbeiter vor das Dilemma, dass eine Tiefenerschließung eines Kataloges mit einem erheblichen Aufwand verbunden ist. Auch wenn Kataloge serielle Quellen und formalisierte Sachtexte sind, die ihre bibliographischen Informationen repetitiv, nach demselben Muster aneinanderreihen, so bedeutet das nicht, dass die Aufnahmen einer einheitlichen Syntax folgen. Es werden die bibliographischen Informationen in Katalogen nicht selten fragmentiert oder bis zur Unkenntlichkeit reduziert wiedergegeben. Die Titelaufnahmen der Frühen Neuzeit gehorchen freilich nicht modernen Ansetzungsregeln, und selbst wenn Titelaufnahmen nach einem vorliegenden Titelblatt eins zu eins abgeschrieben werden, gibt es als Gegensatz dazu auch Einträge, die nicht den Titel nennen, sondern den Inhalt des Buches paraphrasieren. So werden die Autorennamen und Buchtitel häufig abgekürzt, es fehlen Druckorte oder/und Druckfehler erschweren die Identifikation.

Für die geisteswissenschaftliche Forschung ist indes seit langem klar, dass sich die Beschäftigung mit Auktionskatalogen auszahlt – eben weil die durch sie rekonstruierbaren Bibliotheken immer mehr erzählen als ihre Bücher (Raabe 1984, passim). Waren es bisher vor allem Untersuchungen zu Einzelkatalogen oder kleineren Korpora, sollte es im Zuge der Etablierung der DH mittelfristig dazu kommen, Mechanismen bereitzustellen, die quantitative Auswertung von Auktionskatalogen auf der

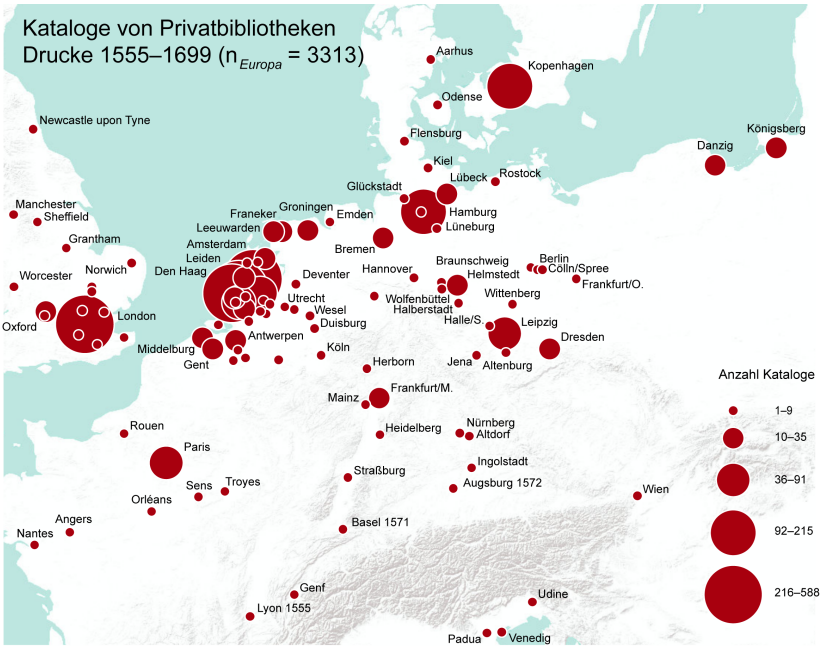


Abbildung 9: Die Häufigkeit gedruckter Kataloge von Privatbibliotheken in Europa nach Erscheinungsorten (Kartenausschnitt Mitteleuropa). Die Daten wurden mit Angaben aus Gerhard Loh 1995ff. und 1997ff. (vgl. Bibliographie) ergänzt; Grafik u. Screenshot: Dietrich Hakelberg

Basis von mit Normdaten und kontrollierten Vokabularen angereicherten Volltexten und Metadaten zu erlauben. Darauf aufbauend können computergestützte Verfahren wie Distant Reading, Topic Modeling oder Netzwerkanalysen hilfreich sein. Ferner kann durch die Erschließung des Quellentyps ›Auktionskatalog‹ die Erforschung von bio-bibliographischen Netzwerken unterstützt werden. Die Anzahl der überlieferten gedruckten Verkaufskataloge von Privatbibliotheken veranschaulicht bereits ihr Quellenpotential für diese Arten einer quantitativen Auswertung (Abb. 9).

Im Rahmen des MWW-Projektes sollte keine Korpusbildung vorgenommen werden, vielmehr ging es um die Entwicklung exemplarischer Erschließungsmodelle, die u.a. ausgewählte Auktionskataloge erschließen helfen. Das Projektteam identifizierte folgende Ziele für die Anwendung von DH-Methoden:

- Die quantitative und inhaltliche Vergleichbarkeit der untersuchten Bibliotheksbestände und die Nachnutzbarkeit der erhobenen Daten ist zu gewährleisten.
- Die Darstellung von ausgewählten frühneuzeitlichen Gelehrtenbibliotheken erfolgt im historischen Kontext.

- Es ist die Visualisierung von Kontexten und Relationen angestrebt.
- Ebenso steht die Präsentation der Forschungs- und Erschließungsergebnisse im Internet auf dem Programm.

Die Basis für die Umsetzung dieser Punkte besteht in der Erschließung des Katalogmaterials über deskriptive Metadaten und die Normierung der enthaltenen Angaben. Flankiert wird dieser Arbeitsschritt von einer Vernetzung zu relevanten Wissensressourcen. Dafür wird der Volltext der bibliographischen Einträge zwingend benötigt. Aufgrund der häufig schlechten Papierqualität, des Widerdrucks, der Schrift- und Sprachwechsel liefert leider keine bekannte OCR-Software einen auch nur ansatzweise wissenschaftlich verwertbaren Volltext der selektierten Katalogstücke. Aufgrund der schon angesprochenen Heterogenität des Quellenmaterials würde auch ein zeitintensives Training der OCR-Software nur dann vertretbare Ergebnisse erzielen, wenn mehrere Auktionskataloge aus derselben Druckerwerkstatt bzw. aus demselben Druckstock stammen. Es bleibt zu hoffen, dass diese Problematik in Zukunft gelöst werden kann, entsprechende Initiativen existieren bereits, wie beispielsweise die *Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR).*[6] Selbst wenn ein wissenschaftlich nachnutzbarer Volltext automatisiert erstellt werden könnte, wäre dieser aufgrund der vielen Abkürzungen oder sogar fehlenden bibliographischen Informationen nur schwer zugänglich und würde sich gegen eine belastbare Auswertung, die auf automatisierten Suchen oder Abfragen beruht, sperren. Das Projektteam entschied sich daher, die Auktionskataloge zu transkribieren und über Metadaten und eine umfangreiche Normierung zu erschließen. Diese Kombination zweier Methoden schafft nicht nur einen stabilen Text, sondern stellt der in Entwicklung begriffenen Webpräsenz auch Daten für die Inkorporation und Analyse zur Verfügung. Die in Excel erhobenen Daten werden im Laufe des Projektes auch in Form von XML/TEI-P5 zur Verfügung gestellt und können daher problemlos von anderen Forschungsprojekten oder Einzelforschern nachgenutzt werden. Der Schwerpunkt der Kodierung liegt bei Auktionskatalogen in der Auszeichnung von bio-bibliographischen Informationen. Da dies ein Bereich ist, der durch die Regeln der TEI detailliert abgedeckt ist, können die Inhalte und Strukturen der Auktionskataloge in toto ausgezeichnet werden.

Der nächste Arbeitsschritt bestand in der Profilbildung: Wie sind die Auktionskataloge zu erschließen und welche weiteren Prozesse sollen der Erschließungsarbeit nachfolgen? Dazu wurden zwei Verfahrensprofile entwickelt. Vor der Zuordnung eines Katalogs zum entsprechenden Profil, muss jedoch eine stichprobenhafte Teilerschließung durchgeführt worden sein. Das Profil *Gelehrtenbibliothek 1* betrifft

[6] In der Schwierigkeit der automatisierten Volltexterkennung (OCR) sind Auktionskatalog mit der Textsorte der frühneuzeitlichen Leichenpredigten durchaus vergleichbar (Federbusch und Polzin 2013, passim).

Auktionskataloge, bei denen der historische Buchbestand rekonstruiert werden konnte. Die Überlieferungssituation, z. B. durch Provenienzeinträge, und die Nachweise im Katalog sind so gut, dass die Granularität den Exemplarvergleich gestattet. In dem Fall ist es sogar möglich, den verstreuten Buchbestand wie im *Virtuellen Skriptorium St. Matthias* virtuell wieder zu vereinigen.

Das Profil *Gelehrtenbibliothek 2* bietet diese Möglichkeit nicht, die Gründe dafür liegen häufig in rudimentären bibliographischen Angaben. Daher findet die Identifikation nicht auf Exemplarebene statt, den gelisteten Titeln in den Verkaufskatalogen kann man sich nur über die Ausgabe- bzw. Werkebene annähern. Hier werden, wie im ersten Fall, der Katalog und die Einträge transkribiert und zusätzlich ein digitales Faksimile des Kataloges zur Verfügung gestellt. Beide Profile verfügen über einen einleitenden Text und über eine Liste von Quellen, die den Katalog flankieren. Nach dem Arbeitsschritt der Inhaltstranskription werden die Daten mit Normdaten angereichert, womit sie für unterschiedliche Visualisierung- und Präsentationsformen zur Verfügung stehen. Nachdem dieser Workflow konzipiert und etabliert worden ist und die relevanten Normdaten und Tools für die Visualisierungen identifiziert und eingerichtet worden sind, kann die Datenaufnahme und Verarbeitung der Kataloge durchgeführt werden. Als gewinnbringend hat sich auch erwiesen, den Workflow nicht als statisches Konstrukt zu begreifen, sondern ihn auf ursprünglich nicht bedachte Erschließungsphänomene anzupassen. Dazu war der regelmäßige Austausch von DHler und Bestandserschließer-Forscher notwendig und förderlich.

Als bereits umgesetzten Fall für das skizzierte Vorgehen im Profil *Gelehrtenbibliothek 1* kann die Edition des Bücherinventars der Elisabeth von Calenberg (1510–1558) (Bücherinventar Calenberg) dienen. Diese Webpräsenz kombiniert editorische Bestandteile mit einer über Normdaten und Metadaten erfolgten Erschließung des Bücherinventars der Herzogin. Stellt dieses Beispiel einen Einzelfall dar, so können der beschriebene Workflow und die erzielten Ergebnisse bei den Auktionskatalogen, wie z. B. im Katalog von Bahnsen, hoffentlich dazu führen, dass der Forschung bald ein relevantes Korpus an frühneuzeitlichen Auktionskatalogen zur Verfügung steht.

In den skizzierten Workflow wurden auch Quellen aufgenommen, die die Auktionskataloge flankieren und einer fundierten Auswertung der primär über die Kataloge fassbaren Büchersammlungen und der hinter ihnen liegenden Gelehrtenbiographien zuarbeiten. In den Kontext der Auseinandersetzung mit dem Verkaufskatalog der Bücherei Bahnsen gehört bspw. das Buchagententum in der Frühneuzeit. Herzog August der Jüngere von Braunschweig-Lüneburg (1579–1666) war für die Vermehrung und Qualität seines Buchbestands maßgeblich auf bibliographisch kompetente Agenten und Informanten angewiesen. Sie, die selbst geschäftliche und persönliche Netzwerke unterhielten oder Teil von ihnen waren, unterrichteten den Fürsten aus erster oder zweiter Hand aus deutschen und europäischen Städten wie Augsburg, Nürnberg, Paris, Rom, Venedig, Den Haag und Amsterdam über den Buchmarkt und

die Angebote. Daneben unterstützten sie ihren Auftraggeber bei der allgemeinen Nachrichtenrecherche und übernahmen zum Teil auch diplomatische Aufgaben an den auswärtigen Höfen. Als Gegenleistung erhielten sie zuzüglich der verausgabten Beträge und Spesen ein festes Salär oder eine vereinbarte Vergütung. Die Agenten konnten zudem auf den Zuwachs von Prestige und gute Absatzmöglichkeiten anderer exquisiter Kulturprodukte hoffen, die sie ebenfalls vertrieben, weil der Kontakt zum Hochadel in der Regel Ansehen und eine zahlungskräftige Kundschaft bedeutete (Arnold 2014a, 81-86; Arnold 2014b, 16-19, 22-25). Neben dem in Nürnberg ansässigen, den Büchererwerb weit über Süddeutschland ausdehnenden Philipp Hainhofer und dem in Paris tätigen und von dort berichtenden Jean Beeck – um nur die beiden bekanntesten zu nennen – gehört auch der hier im Mittelpunkt stehende Benedikt Bahnsen zu den Buch- und Geschäftsagenten, die für Herzog August Dienst taten. Wir wissen nicht, ab wann genau Bahnsen in den Niederlanden für den berühmten Wolfenbütteler Büchersammler Drucke und Handschriften erwarb. Es ist davon auszugehen, dass er spätestens seit dem Frühjahr 1660 in entsprechendem Einsatz war. Diese Annahme gründet auf der erhaltenen Korrespondenz zwischen Bahnsen und Herzog August, die sich in der Herzog August Bibliothek in Wolfenbüttel befindet (HAB Wolfenbüttel: BA II,1, Nr. 16–35). Der Briefwechsel umfasst 21 Briefe Bahnsens an den hochgelehrten Fürsten, mehrere Bücherlisten und Rechnungen sowie drei Konzepte von dessen Antwortschreiben aus der Zeit zwischen April 1660 und Juli 1666, also bis kurz vor Herzog Augusts Tod im September 1666. Aus dem Jahr 1664 liegen keine Briefe vor. Da Bahnsen über Ware und Aufwand Rechnungen stellte, wird er – wie Dietrich Hakelberg plausibel schlussfolgert – wohl keine hervorgehobene Rolle als Bücheragent gehabt und auch kein festes Salär bezogen haben. Ebenso wenig dürfte er mit diplomatischen Aufgaben betraut gewesen sein (Hakelberg 2015b, 136).

Der Briefwechsel setzt mit einem Schreiben Bahnsens aus Amsterdam an den Fürsten in Wolfenbüttel am 4. April 1660 ein (Abb. 10). Dieses Datum markiert aber nicht den Auftakt des Kontaktes zwischen den beiden Männern, da Bahnsen in dem Brief auf eine vorgängige Anfrage von Herzog August reagiert, deren Wortlaut aber nicht erhalten ist. Das Schreiben kann als Muster für die gesamte Korrespondenz gelten, insofern es bis auf eine mögliche Titelliste, ein nachgestelltes Bücherverzeichnis und/oder eine Beilage von anderer Hand im Anhang die Hauptbestandteile der epistolaren Meldungen von Bahnsen enthält: Datum und Ort (stets Amsterdam), die formelhafte Anrede des Fürsten, Bezugnahme auf bestellte Bücher, ihre Nicht-/Verfügbarkeit, Frachtwege, Erwähnung von an der Bücherbeschaffung und der Bücherspedition beteiligten Personen, Preise, Spesen, sonstige Aufwendungen, formelhafte Verabschiedung, schließlich der Hinweis auf anhängende Bücherpakete. Natürlich sind für uns die gelisteten Titel, Autorennamen und Hinweise auf tatsächliche Büchersendungen an Herzog August von größtem Interesse. Hinzu kommen Aussagen, die auf Verbindungen, gar Netzwerke mit anderen Gelehrten, Bücherproduzenten, -händlern und

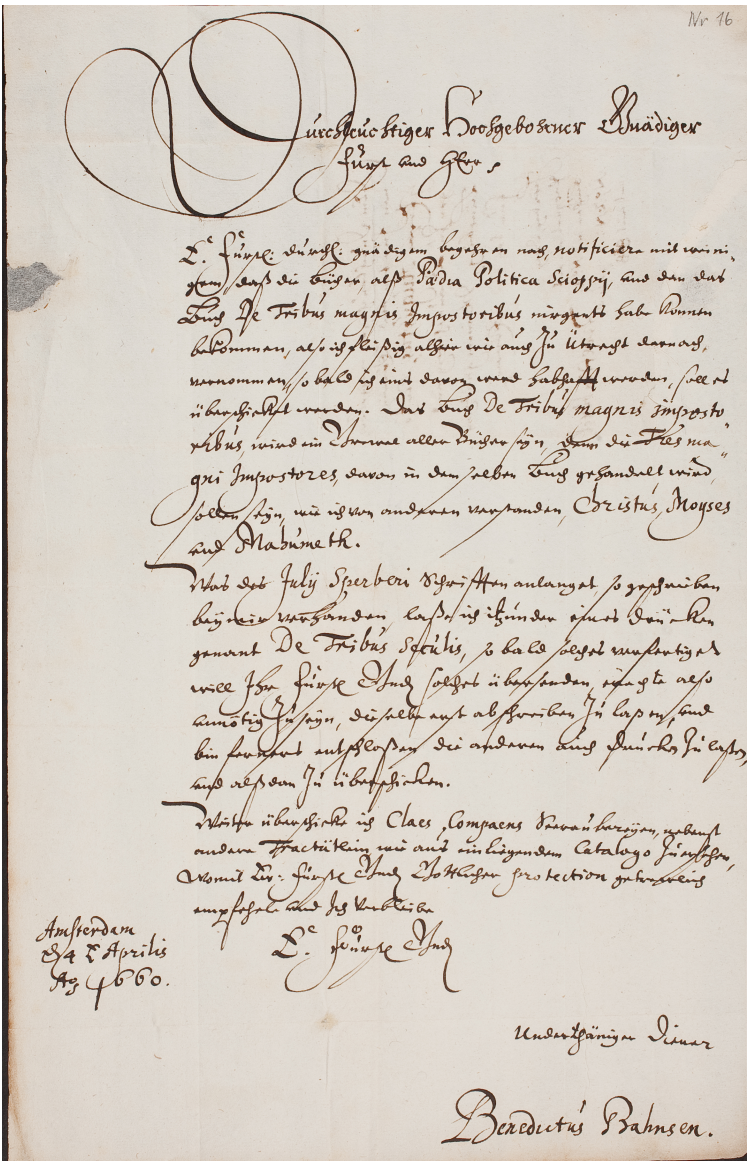


Abbildung 10: Brief von Benedikt Bahnsen an Herzog August d.J. v. Braunschweig-Lüneburg (4. April 1660, Amsterdam), Sign.: HAB, Bibliotheksarchiv, BA II,1, Nr. 16, 1r.

weiteren historisch eminenten Personen verweisen. Ebenso spielen Auskünfte zum genauen Wohnort Bahnsens eine Rolle, zu seinen Beobachtungen und Beurteilungen zeitgenössischer Geschehnisse, um so seine historische Person schärfer zu fassen und ein möglichst umfassendes Profil zu rekonstruieren. Es können bibliophile Vorlieben Bahnsens, wie sie am Verkaufskatalog seiner Bücherei ablesbar wurden, verifiziert oder korrigiert werden, es lassen sich über die individuelle Stimme, die Angebote, Besorgungsversprechungen, Vollzugsmeldungen und Klagen in den Briefen zudem seine genauen Lebensumstände präzisieren. Das Briefkorpus gehört, das abschließend dazu, wesentlich zum Primärquellenmaterial, seine Analyse ermöglicht einen mikroperspektivischen Blick auf die Existenz des frühneuzeitlichen Buchhändlers und Verlegers, Chiliasten, Exilanten und Befürworters heterodoxer bzw. religiös devianter Literatur, Rechenmeisters und Gelehrten Benedikt Bahnsen.

In der geplanten digitalen Edition dieser Briefe erfolgt die Auszeichnung in XML/TEI-P5. Ausgezeichnet werden im epistolaren Text zum einen briefrelevante Informationen wie Adressat, Grußformel, Schreibdatum etc. Zum anderen sollen die genannten Personen, Orte und vor allem Buch- und Werktitel, also die bibliographischen Titelangaben über Normdaten erschlossen werden. Die Edition der Briefe wird mit dem Datenertrag aus dem Bahnsen-Auktionskatalog und der von Herzog August bis zum seinem Tod 1666 selbst geleisteten bzw. unter seiner Aufsicht bibliographisch erfassten Wolfenbütteler Sammlung verknüpft. So soll möglich sein herauszufinden, welche von Herzog August bei Bahnsen angeforderten Titel tatsächlich von diesem beschafft und nach Wolfenbüttel versendet wurden und sich noch heute im ›Urbestand‹ der HAB identifizieren lassen. Die Briefe werden als Editionspaket in der *Wolfenbütteler Digitalen Bibliothek (WDB)* publiziert und können im Volltext durchsucht werden. Ferner werden die Briefe auch katalogisiert und bieten durch ihren Nachweis in Repositorien zu edierten Briefen (z.B. das Verzeichnis zu Briefeditionen *CorrespSearch*) einen zusätzlichen Einstieg zu dem konkreten Auktionskatalog.

Auch wenn die Bahnsen-Erschließung nur als Usecase konzipiert ist, wird dennoch ersichtlich, dass Auktionskataloge mittels noch profunder in ihrer historischen Bedingtheit erfasst, gewichtet und ausgelotet werden können. Das Netzwerk zwischen Personen und Bibliographien lässt sich durch weitere Referenzen (Quellen und Normdaten) detailliert erforschen und darstellen. Dies kann außerdem den Grundstein dafür legen, in Zukunft auch für einzelne Bücher eine Art ›Biographie‹ rekonstruieren zu können. Die sich erst im Laufe des Bahnsen-Projektes ergebene und entwickelte Flankierung der Auktionskataloge mit einer Briefedition wäre in einem analog ausgerichteten Projekt nicht ohne weiteres möglich gewesen und demonstriert die erweiterten Möglichkeiten, die sich der Forschung durch die DH eröffnen.

Bibliographie

Adam, Wolfgang. »Bibliotheksforschung als literaturwissenschaftliche Disziplin.« In Hölter, Achim und Stefan Alker (Hrsg.). *Literaturwissenschaft und Bibliotheken*. Göttingen: V & R unipress GmbH, 2015. 67-92.

Arnold, Werner [2014a]. »Philipp Hainhofer als Bücheragent Herzog Augusts d.J. von Braunschweig-Wolfenbüttel.« *Wolfenbütteler Barock-Nachrichten*, 41.1/2. (2014). 81-94.

Arnold, Werner [2014b]. »Briefe aus Paris. Jean Beeck als Agent Herzog Augusts d.J. von Braunschweig-Wolfenbüttel.« *Buch – Bibliothek – Region. Wolfgang Schmitz zum 65. Geburtstag*. In Haug, Christine und Rolf Thiele (Hrsg.). Wiesbaden: Harrassowitz Verlag, 2014. 15-26.

Ball, Gabriele. »Privatbibliotheken – eine Einführung.« In Schneider, Ulrich Johannes (Hg.). *Kulturen des Wissens im 18. Jahrhundert*. Berlin/New York (NY): De Gruyter, 2008. 191-194.

Bücherinventar der Elisabeth von Calenberg (1510–1558), hg. v. Eva Schlotheuber und Gabriele Haug-Moritz, unter Mitarb. von Anna Durwen, Eva Glaser und Stephanie Moisi. Herzog August Bibliothek Wolfenbüttel, 2011. <http://diglib.hab.de/edoc/ed000082/start.htm>.

CorrespSearch. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. <http://correspsearch.net/>.

Coy, Wolfgang. »Gutenberg & Turing. Fünf Thesen zur Geburt der Hypermedien.« *Zeitschrift für Semiotik* 16.1/2 (1994). 69-74.

Cruz, Laura. *»Auctions.« The Paradox of Prosperity. The Leiden Booksellers' Guild and the Distribution of Books in Early Modern Europe*. New Castle (DE): Oak Knoll Press, 2009. 103-146.

DARIAH Geo-Browser und Datasheet Editor. Göttingen: Niedersächsische Staats- und Universitätsbibliothek. <https://de.dariah.eu/geobrowser>.

Federbusch, Maria und Christan Polzin. *Volltext via OCR – Möglichkeiten und Grenzen. Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz*. Berlin: Staatsbibliothek zu Berlin, 2013.

MWW: *Forschungsverbund Marbach Weimar Wolfenbüttel*. 2013-. <http://www.mww-forschung.de>.

Hakelberg, Dietrich [2015a]. »Die Bücherschenkung des Augsburger Patriziers Carl Wolfgang Rehlinger von 1575 und ihr gedruckter Katalog.« In *Gutenberg-Jahrbuch* 90 (2015). 216-234.

Hakelberg, Dietrich [2015b]. »Die fanatischen Bücher des Benedikt Bahnsen. Leben und Bibliothek eines religiösen Dissidenten.« In Knoche, Michael (Hg.). *Autorenbibliotheken. Erschließung, Rekonstruktion, Wissensordnung*. Wiesbaden: Harrassowitz, 2015. 113-146.

Kaden, Ben. »Wer übernimmt was? Zum Verhältnis von Digital Humanities und Geisteswissenschaften.« *LIBREAS. Debatte*, 12. September 2013. <https://libreas.wordpress.com/2013/09/12/digital_humanities/>.

Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren für die Optical-Character-Recognition (OCR), Wolfenbüttel: Herzog August Bibliothek, Berlin: BBAW, Staatsbibliothek Preußischer Kulturbesitz. 2015-2018. <http://www.ocr-d.de/>.

Loh, Gerhard. »Verzeichnis der Kataloge von Buchauktionen und Privatbibliotheken aus dem deutschsprachigen Raum.« Leipzig: Loh [o.V.], 1995ff.

Loh, Gerhard. »Die europäischen Privatbibliotheken und Buchauktionen. Ein Verzeichnis ihrer Kataloge.« Leipzig: Loh [o.V.], 1997 ff.

Neuroth, Heike et al. »Virtuelle Forschungsumgebungen für e-Humanities. Maßnahmen zur optimalen Unterstützung von Forschungsprozessen in den Geisteswissenschaften.« In *BIBLIOTHEK Forschung und Praxis*, 33.2 (2009). 161-169.

Pozzo, Annette. *Membra disiecta. Inhalt und Wirkung der Bibliothek des Göttinger Professors Lüder Kulenkamp (1724-1794)*. Berlin: Humboldt Universität zu Berlin, Diss., 2013. URN: urn:nbn:de:kobv:11-100213427.

Privatbibliothek des Mathematikers, Astronomen und Arztes Duncan Liddel (1561–1613), hg. v. Karin Friedrich et al., <http://uni-helmstedt.hab.de/index.php?cPage=8&sPage=liddel>.

Raabe, Paul. »Bibliothekskataloge als buchgeschichtliche Quellen. Bemerkungen über gedruckte Kataloge öffentlicher Bibliotheken in der frühen Neuzeit.« In Wittmann, Reinhard (Hg.). *Bücherkataloge als buchgeschichtliche Quellen in der frühen Neuzeit*. Wiesbaden: Harrassowitz, 1984. 275-297.

Sahle, Patrick. »Digital Humanities? Gibt's doch gar nicht!« In Baum, Constanze und Thomas Stäcker (Hrsg.). *Grenzen und Möglichkeiten der Digital Humanities*. 2015 (= Sonderband der ZfdG 1). DOI: 10.17175/sb001_004.

TEI: *Text Encoding Initiative*. 1987-. <http://www.tei-c.org/>.

Virtuelles Skriptorium St. Matthias, Trier: Trier Center for Digital Humanities, Stadtbibliothek Trier. 2010-2014. <http://stmatthias.uni-trier.de/>.

Vogeler, Georg. »Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?« In Baum, Constanze und Thomas Stäcker (Hrsg.). *Grenzen und Möglichkeiten der Digital Humanities*. 2015 (= Sonderband der ZfdG 1). DOI: 10.17175/sb001_007.

Wieland, Magnus. »Stell-Werk. Literatur im Bücherregal.« In *Autorenbibliotheken / Bibliothèques d'auteurs Quarto: Zeitschrift des Schweizerischen Literaturarchivs*. 30/31 (2010). 27-33.

WDB: *Wolfenbütteler Digitale Bibliothek*. Wolfenbüttel: Herzog August Bibliothek. <http://www.hab.de/de/home/bibliothek/digitale-bibliothek-wdb.html>.

The Legendary Legacy: Crunching 600 Years of Saga Manuscript Data

Matthew Driscoll

Abstract

The research project "Stories for all time", which ran from 2011 to 2014, had as its aim to survey the entire transmission history of the *Fornaldarsögur Norðurlanda*, a group of Icelandic sagas often referred to in English as "mythical-heroic" or "legendary" sagas. Although the sagas themselves are thought to date from the 13th or 14th century, they are preserved mostly in post-medieval paper manuscripts. We set out therefore to locate and catalogue all the manuscripts containing texts of the 35 sagas which make up the corpus. In the end we found over 1000 manuscripts – containing nearly 2000 texts – the earliest from the beginning of the 14th century, the latest from the beginning of the 20th. About a quarter of these were not previously known to scholarship. We catalogued all of these manuscripts using a very restrictive subset of the TEI manuscript description module, which allows us to compare codicological and other features of the manuscripts in a way hitherto impossible. The article presents the schema and some of the results of our analysis of the encoded data.

Zusammenfassung

Im Rahmen des Forschungsprojekts »Stories for all time« wurde in den Jahren 2011-2014 die Überlieferungsgeschichte aller unter dem Namen *Fornaldarsögur Norðurlanda* gefassten und als »mythisch-heroisch« oder »legendenhaft« bezeichneten isländischen Sagen untersucht. Ihre Entstehung wird für gewöhnlich in das 13. und 14. Jahrhundert datiert; überliefert sind sie gleichwohl vor allem in neuzeitlichen Papierhandschriften. Ziel des Projekts war es, sämtliche Textzeugen des 35 Sagen umfassenden Corpus aufzufinden und zu katalogisieren. 1000 Handschriften mit etwa 2000 Texten konnten identifiziert werden, die älteste vom frühen 14. Jahrhundert, die jüngste vom frühen 20. Jahrhundert. Ein Viertel aller Textzeugen waren der Forschung zuvor noch unbekannt. Die Handschriften wurden unter Verwendung eines sehr strikten TEI-Schemas katalogisiert, das einen bis dato nicht möglichen Vergleich kodikologischer und anderer Eigenschaften erlaubt. Dieser Artikel stellt sowohl das Schema selbst als auch die Ergebnisse einer Analyse der mit diesem Schema erfassten Daten vor.

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 71–79.

The project "Stories for all time: The Icelandic fornaldarsögur", based at the University of Copenhagen, has as its chief focus the transmission history of the *Fornaldarsögur Norðurlanda* – literally "Stories of the ancient men of the northern lands" but generally known in English as Legendary or Mythical-heroic sagas – a group of Icelandic prose narratives dealing with the early history of mainland Scandinavia, before the unification of Norway under Haraldr *hárfagri* (fair-hair) and the settlement of Iceland in the late 9th century. Although many of them demonstrably have older roots, the sagas as we have them were first written down in the 14th century. They remained popular throughout the late medieval and early-modern period, even into the 18th and 19th centuries and the first decades of the 20th.[1]

The project's chief deliverable is an electronic catalogue of all the manuscripts in which *fornaldarsaga* texts are found, containing information on their format and layout, the other texts they preserve and when, where and by and/or for whom they were written. Ancillary to this is a fully searchable bibliography of editions, translations and secondary material pertaining to the *fornaldarsögur*. Both the manuscript catalogue and the bibliography were produced in TEI-conformant XML. Both are regularly updated and available on the project website.

So far, 817 manuscripts have been identified as containing *fornaldarsaga* texts; about a quarter of these were not previously known to scholarship.[2] Of these, 82 are composite manuscripts, i.e. are made up of parts (two or more) of originally separate manuscripts. If the parts are counted separately, the total number of manuscripts is 1049 (a typical Fornaldarsaga manuscript is shown in fig. 1).

Most are from Iceland, but some were written, generally by or for scholars, in Sweden or Denmark. And although most are in Icelandic, about 150 are, or contain alongside the Icelandic text, translations into Swedish, Danish, French or Latin. Only around a quarter of the manuscripts only contain *fornaldarsögur*; the rest contain material belonging to other genres, principally *riddarasögur* (chivalric romances, both translated and original) and *Íslendingasögur* (Icelandic family sagas), but all sorts of other things as well (see further below).

For each manuscript there is a catalogue record produced using a restrictive subset of the TEI P5 module for manuscript description.[3] Among other things, the number of elements available for use was greatly reduced, many elements and attributes which are optional in the TEI were made mandatory, and many attribute value lists were 'hard-wired' into the schema. This was done both to make data-input easier for the cataloguers and reduce the risk of error, and also to make the data more easily searchable. We have for the same reason deliberately tried to put as much information

1 For a definition of the *fornaldarsögur*, see Driscoll 2003 and Driscoll 2009.

2 This number will certainly increase as more manuscripts in private ownership are discovered and catalogued.

3 The module for manuscript description is presented in chapter 10 of the *TEI Guidelines*.

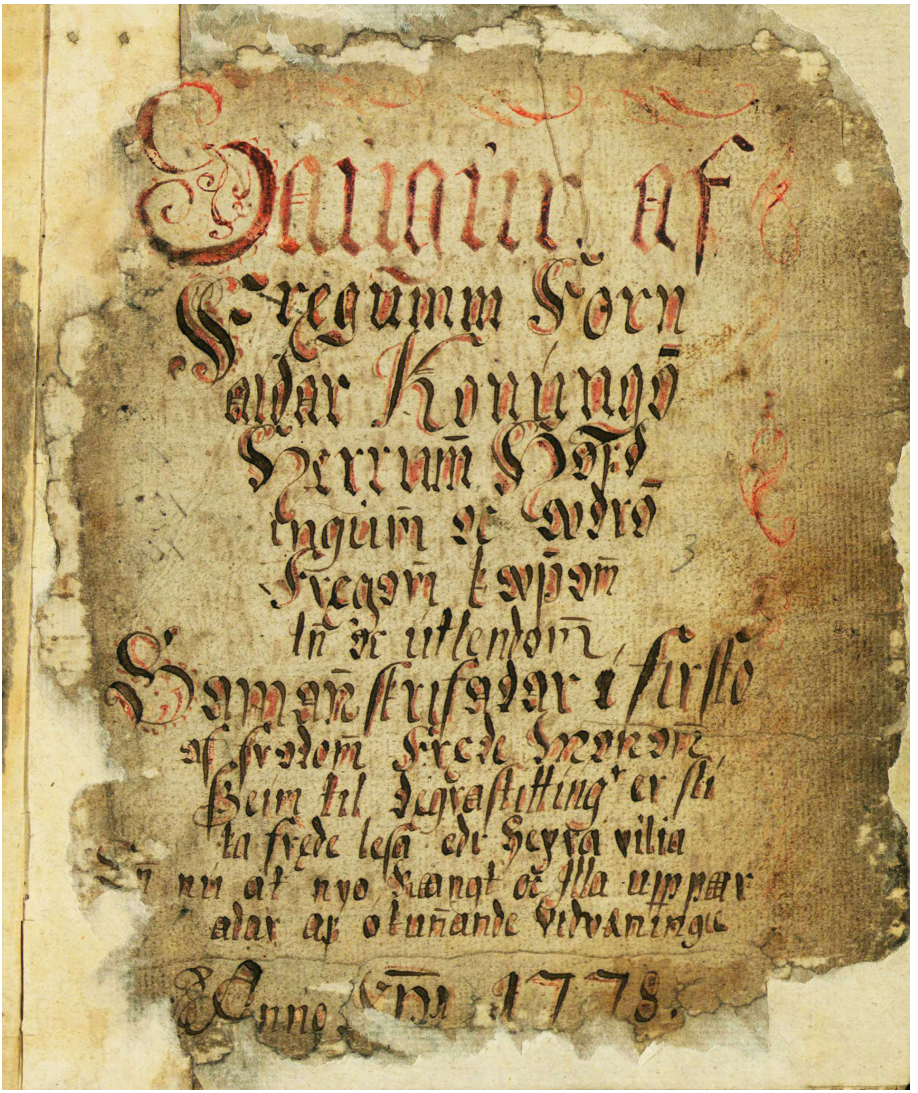


Figure 1: Reykjavík, Landsbókasafn Íslands – Háskólabókasafn, ÍB 165 4to, a large collection of Apostles' *vitæ* and *fornaldarsögur* written in 1778 in Selárdalur by an unknown scribe, who identifies himself only as "P. J.son". According to the title-page, shown here, the sagas were "Saman*n* skrifadar í firstu af fródum*m* fręde mønnum*m*, þeim til dęgrastittíng*ar* er slíka frędę lesa edr heyra vilia, en nú at nyo rangt oc jlla upparadar af ókun*n*ande vidvaningie" (originally compiled by wise men of learning for the enjoyment of those who wish to read or hear such lore, but now once again badly and inaccurately scrawled by an ignorant amateur).

into the encoding as we can, avoiding wherever possible the use of natural language. So instead of indicating the language of the text by using the word "Icelandic" or "Swedish", for example, like this:

```
<textLang>Icelandic, with some Swedish</textLang>
```

we would put this information in attribute values, like this:

```
<textLang mainLang="is" otherLangs="sv"/>
```

The way this is displayed in the online database is then determined by the stylesheet. One added advantage of this method is that content can then be generated in any language, should one want to have the option of multiple interface languages.

As in standard TEI, the `<msDesc>` (manuscript description) element contains a description of a single identifiable manuscript. In our schema it must have the attributes `@xml:id`, which provides a unique identifier for the element, and `@xml:lang`, which indicates the language of the element content. The sub-elements of `<msDesc>` are `<msIdentifier>`, `<msContents>`, `<physDesc>`, `<history>` and `<additional>`, all of which should be present. Two further elements, `<msPart>` and `<msFrag>`, are also available within `<msDesc>`; the former is used for composite manuscripts, i.e. manuscripts comprising two or more originally distinct manuscript parts now kept together as a unit, and the latter for scattered manuscripts, i.e. manuscripts one or more parts of which have become separated from the original codex and may now be kept in different repositories.

Each of the child elements of `<msDesc>` contains a number of sub-elements, many of which have also been made mandatory. `<msContents>`, for example, must contain at least one `<msItem>` element, on which the attributes `@class` and `@n` must be present. Each `<msItem>`, in turn, must contain the elements `<locus>`, `<title>` and `<textLang>`, each with their required attributes.

```
<msContents>
 <msItem class="#fas" n="1">
 <locus from="1r:1" to="8v:17"/>
 <title type="uniform" ref="#snfdsv">
  Sögubrot af nokkrum fornkonungum í Dana ok Svía veldi
 </title>
 <textLang mainLang="is"/>
 </msItem>
</msContents>
```

One recurrent topic of debate within *fornaldarsaga* studies has been that of genre: to which extent can or should they be considered to represent a corpus?[4] Apart from the criteria of the time and geographical space in which the stories are set, do they share any features which may be said specifically to characterise them and distinguish them from other types of sagas. And, more importantly, is there any evidence in the

[4] One of the more recent contributions to this debate is Quinn et al. 2006.

manuscripts themselves to suggest that those who copied and read them regarded them as constituting a genre?

In order better to address this question we have attempted to identify the other types of texts found in manuscripts alongside the *fornaldarsögur*, which is why the `@class` attribute has been made mandatory on all `<msItem>` elements. The possible values for `@class` are defined in a `<taxonomy>` element in the TEIheader. The different class designations are based on the indexing terms used by the National Library of Iceland, but simplified greatly.

We also place a lot of emphasis on the manuscripts' codicological features, and so many of the elements within `<physDesc>` (physical description) are also mandatory. Describing such features can be very time-consuming, however, and since we had a large number of manuscripts to get through in a relatively short period we developed a simple 'short cut' which allows us to provide basic information on the presence or absence of a feature or its relative level or extent without the necessity of going into any further detail. Flagging the presence of a feature in this way allows us to return to the manuscript later if need be. To this end the attribute `@ana` (analysis) is used on a number of elements.

To take one example: title pages, which were not a feature of medieval manuscripts but developed after the invention of moveable type, are often found in younger, post-medieval, manuscripts. In order simply to register their presence, and whether they appeared to be contemporary with the manuscript or added later, we require the attribute `@ana` on the element `<titlePage>`, with possible values "no", "contemporary", "later" and "unknown". No further content is required, but sub-elements such as `<titlePart>` can be used, or added later.

Other elements which can (or must) also use `@ana` in this way include `<foliation>`, `<watermark>`, `<condition>`, `<decoDesc>` and `<additions>`. In the latter two cases the possible values are "no", "low", "medium" and "high"; no other information need be supplied.

It could be argued that this is misuse of the `@ana` attribute, which is intended to provide a pointer "to one or more elements containing interpretations of the element on which the `@ana` attribute appears",[5] and that if a manuscript contains no watermarks, say, the best way to indicate this is by simply not using the `<watermark>` element. We disagree, however; the absence of an element does not necessarily indicate the absence of the feature that element is intended to be used to describe.[6]

---

[5] TEI Guidelines, section 17.2, "Global Attributes for Simple Analyses".

[6] This matter was discussed, though with no conclusion being reached, on the TEI listserv in February 2010; see <http://tei-l.970651.n3.nabble.com/Re-Indicating-the-presence-or-absence-of-a-feature-td2349886.html#a2349891>.

Within <support> we use <num> (number) to indicate the number of the leaves, and <dimensions> to indicate their size (a visualisation of leaf-size over time is given in fig. 2). As mentioned above we try to put as much information into the encoding as we can, as in the following example:

```
<supportDesc material="chart">
 <support>
  <num type="front-flyleaf" value="2"/>
  <num type="book-block" value="19"/>
  <num type="back-flyleaf" value="1"/>
  <dimensions type="leaf" unit="mm" scope="all">
   <height quantity="305"/>
   <width quantity="190"/>
  </dimensions>
 </support>
 <watermark ana="#yes"/>
 <foliation ana="#later #fol"/>
 <condition ana="#average"/>
</supportDesc>
```

The description of the layout is similar, again using <num> to indicate the number of words per line and <dimensions> to indicate the size of the written area:

```
<layoutDesc>
 <layout columns="1" writtenLines="28">
  <num type="wpl" atLeast="10" atMost="12"/>
  <dimensions type="written" unit="mm" scope="all">
   <height quantity="230"/>
   <width quantity="175"/>
  </dimensions>
 </layout>
</layoutDesc>
```

On the basis of this, one can easily work out the density of the text on the page. The proportion of the page taken up by the writing, the 'text-page ratio', can be determined by simply dividing the size of the written area (height × width) by the size of the leaf (height × width). In the case of the manuscript described here this would be 69.5%. A simple way of determining text density is to divide the size of the written area (height × width) by the number of words on the page (no. of lines × no. of words per line), which gives you the average amount of space (in mm^2) taken up by a single word; in this case the figure would be 130.68. The smaller the number, the greater the text density. There are, of course, other ways to measure text density, for example by the average amount of space taken up by a single sign (whether letter, abbreviation or mark of punctuation), or the number of signs per unit of space, typically dm^2.[7] Both of these are more time-consuming than the method outlined here, which, despite its 'quick and dirty' nature, does give a good indication of the density of the text on the page which can be used as a basis for quantitative analysis.

[7] See Maniaci 2002, esp. 101-120, and Gumbert 2010, 50-53. There are also software programs which can measure the relative amounts of ink and white space on a page and thus measure the density of the text; see e.g. Gurrado 2009.

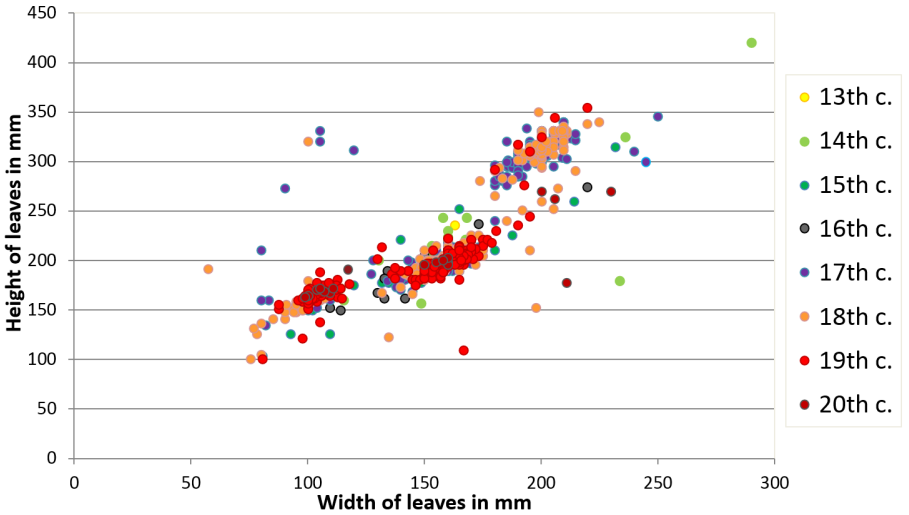


Figure 2: Manuscript leaf-size over time. Visualisation by Beeke Stegmann.

In addition to information on the manuscript's support and layout, the schema allows data on other characteristics to be supplied in a similarly data-intensive fashion. These include:

- number of hands in the manuscript and relative scope of each;
- the names of the scribes identified as corresponding to hands in the manuscript;
- the relative level of decoration of the manuscript;
- the relative level of additions (marginalia) made to the manuscript;
- the degree to which the binding is decorated and the contemporaneousness of the binding with the manuscript;
- the date and place of origin;
- the names of previous owners or other individuals known to have had a part in the manuscript's history.

We have also produced authority files for persons and places named in the manuscript descriptions using the `<person>` and `<place>` elements. Here, for example, is the `<person>` element for the scribe Jón Erlendsson:

```
<person sex="1" role="scribe" xml:id="JonErl001">
 <persName xml:lang="is">
  <forename sort="1">Jón</forename>
  <surname sort="2">Erlendsson</surname>
 </persName>
 <birth notBefore="1600" notAfter="1610">ca. 1605</birth>
 <death when="1672-08">August 1672</death>
```

```
 <residence>
  <placeName>
   <settlement type="farm" ref="#VilVil01">Villingaholt</settlement>
  </placeName>
 </residence>
 <occupation xml:lang="en">Priest</occupation>
 <bibl>
  <ref target="#IsAev">Íslenzkar æviskrár</ref>
  <biblScope unit="volume">III</biblScope>
  <biblScope unit="page" from="195" to="106"/>
 </bibl>
</person>
```

This is then referenced in the individual manuscript records like this:

```
<name ref="#JonErl001" type="person">Jón Erlendsson</name>
```

Or within the <handDesc> element like this:

```
<handNote scope="major" scribeRef="#JonErl001" script="textualis">
 Written, apart from fol. 12, by Jón Erlendsson from Villingaholt
 in a clear, seventeenth-century Gothic book hand.
</handNote>
```

Within the `<person>` element itself, the `@ref` attribute on the `<settlement>` element, indicating Jón Erlendsson's place of residence, points to the corresponding `<place>` element in the place name authority file:

```
<place xml:id="VilVil01">
 <placeName xml:lang="is">
  <settlement type="farm">Villingaholt</settlement>
  <region type="parish" ref="#Villin01"/>
  <region type="county" ref="#&#xC1;rnes01"/>
  <region type="geog" ref="#Sunnle01"/>
  <country ref="#IS"/>
 </placeName>
 <location>
  <geo>63.883997 -20.750909</geo>
 </location>
</place>
```

Note that within the `<placeName>` element, all sub-elements, apart from `<settlement>`, which provides the name of the specific place in question, are pointers to other `<place>` elements in the authority.

Although this has not yet been implemented within our project, it would be possible on the basis of this mark-up to generate maps showing manuscript origin; this could help to reveal, among other things, whether certain sagas were more popular in certain places, and whether this changed over time.

All these different types of information can be collated, revealing things like changes in the distribution of texts over time or trends in format and layout. In the graph below, for example, manuscript format is collated with period of writing. The clusters show clearly the three principal formats, folio, quarto and octavo. It is interesting that the 19th-century manuscripts, which were mostly copied by ordinary people

for their own use, tended to be in the smaller formats of octavo and quarto, while those of the 17th and 18th centuries, which were often written by or for scholars and antiquarians, tended to be in folio.

The point of this highly restrictive schema was to allow for the encoding of basic codicological data on a moderately large number of manuscripts, based, wherever possible, on first hand examination of the manuscripts in question. As these manuscripts were held by some 29 repositories in 8 different countries, we were often forced to work at some speed, without the luxury of in-depth inspection. The idea was therefore to make data input as easy as possible, to reduce the possibility of error and to allow the presence or absence of particular features to be recorded, both for statistical purposes, and to flag items potentially of interest for further investigation. Although the resulting electronic catalogue is narrowly focused on one type of late medieval Icelandic narrative, we hope that our schema, or at least our approach, could be used as a model for similar investigations of virtually any body of documents.

Bibliography

Driscoll, Matthew. "Fornaldarsögur Norðurlanda: The stories that wouldn't die." In Jakobsson, Ármann, Annette Lassen, and Agneta Ney (eds). *Fornaldarsagornas struktur och ideologi*. Uppsala: Institionen för nordiske språk, 2003. 257-67.

Driscoll, Matthew. "A new edition of the *fornaldarsögur Norðurlanda*: Some basic questions." In Bampi, Massimiliano and Fulvio Ferrari (eds.). *On editing Old Scandinavian texts: Problems and perspectives*. Trento: Università degli studi di Trento, 2009. 71-84.

Gumbert, J. Peter. *Words for codices: An English codicological terminology*. Lopik: [s.n.], 2010.

Gurrado, Maria. "'Graphoskop', uno strumento informatico per l'analisi paleographica quantitative." In Rehbein, Malte, Patrick Sahle, and Torsten Schaßan. *Kodikologie und Paläographie im digitalen Zeitalter 1 — Codicology and Palaeography in the Digital Age 1*. Norderstedt: BoD, 2009. 251-59.

Fornaldarsögur Norðurlanda: *Stories for all time*. 2011-2014. <http://fasnl.ku.dk>.

Maniaci, Marilena. *Archeologia del manoscritto: Metodi, problem, bibliografia recente*. Roma: Viella, 2002.

TEI Consortium (ed.). "Manuscript Description." In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.0.0. Last updated on 29th March 2016. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html>.

Quinn, Judy et al. "Interrogating genre in the *Fornaldarsögur*: Round-table discussion." *Viking and Medieval Scandinavia* 2 (2006). 276-96.

VisColl: A New Collation Tool for Manuscript Studies

Dot Porter, Alberto Campagnolo, Erin Connelly

Abstract

The principal physical feature of the book in codex format, the gathering structure, is usually not visualized within digitization projects. If this information is recorded at all, it is generally done with the use of collation formulas. There is not a standard schema for manuscript collation formulas and not all practices are able to record accurately the structure of books. There have been some attempts in the past to describe gathering structures in more formalised ways. VisColl is building on past experiences and strives to describe, visualize, and communicate the gathering structure of books. Successful applications of the new tool are presented as examples. Future versions will add functionality to link physical details of a manuscript with additional information about the content, which will enable a complete mapping of a physical manuscript.

Zusammenfassung

Das Hauptmerkmal eines Buchs im Kodexformat, seine Lagenstruktur, wird in Digitalisierungsprojekten gewöhnlich nicht visualisiert. Wenn diesbezügliche Informationen überhaupt festgehalten werden, so geschieht dies in aller Regel unter Verwendung fomalisierter Lagenbeschreibungen, für die es bisher kein allgemein anerkanntes Standardformat gibt. Auch eignen sich vorherrschende Beschreibungspraktiken nicht immer für eine detailgenaue Erfassung der Lagenstruktur. In der Vergangenheit gab es einige Versuche, Lagenbeschreibungen stärker zu fomalisieren. VisColl knüpft an diese Erfahrungen an und ist bestrebt, Lagenstrukturen von Büchern zu beschreiben, zu visualisieren und zu vermitteln. In diesem Artikel soll anhand einiger Beispiele veranschaulicht werden, wie das neue Tool bereits erfolgreich angewendet wird. In Zukunft sollen Funktionalitäten hinzugefügt werden, über die sich Angaben zum materiellen Zustand einer Handschrift mit inhaltlichen Informationen verbinden lassen, um auf diese Weise ein umfassendes Verzeichnen des physischen Objekts zu ermöglichen.

1 Introduction

VisColl is a digital tool designed to help scholars to visualize the physical construction of medieval codex manuscripts, also known as *collation*. Manuscript codices, like

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 81–100.

modern books, consist of a series of pages, however the pages are physically connected in ways that are not always clear to the reader. Manuscripts are built of *quires*, which are normally three to six sheets of parchment or paper (or both), stacked and then folded in half, and then (usually) sewn together in the fold. The folded sheets are called *bifolia* (literally "two folios"), and the pages are called *folios* or *leaves*. Thus the first leaf in a quire is literally half of a bifolia, while the last leaf in a quire is the other half. We say that these two leaves are *conjoined*. Quires are sewn together to create codex books. In addition to sets of bifolia, a quire may have leaves cut out, or added either during the writing process or later. It is these details of physicality—quires, bifolia, added and removed leaves—that the current version of VisColl seeks to describe and visualize. Future versions will add functionality to link physical details of a manuscript with additional information about the content, which will enable a complete mapping of a physical manuscript.

VisColl was conceived in the mid-2000s by Dot Porter during her work at the Collaboratory for Research in Computing for Humanities at the University of Kentucky (UKY). Porter developed the tool in order to address issues she encountered in effectively visualizing standard descriptions of manuscripts in scholarly works. For instance, in *Beowulf and the Beowulf Manuscript* Kevin Kiernan (1981) uses the physical construction of the manuscript to make arguments about the dating of the text (separate from the dating of the manuscript itself). In addition, Ben Withers (of UKY), in *The Illustrated Old English Hexateuch, Cotton MS. Claudius B.IV: the Frontier of Seeing and Reading in Anglo-Saxon England* (2007) similarly used a detailed collation statement of the manuscript as the backbone for his investigation of the construction of the manuscript. There are numerous examples of scholarly works that build an argument about the dating and construction of manuscripts based on the collation of the physical object. In consulting such works, Porter saw an opportunity to enable readers to better visualize the structure of the object beyond the limitations of traditional formulas, diagrams, and collation statements.

1.1 Collation formulas

Traditionally, information on the gathering structure of books is recorded in highly dense expressions, referred to as *collation formulas*. These describe the sequence of bifolia (and singletons) within book gatherings. All formulas contain the same basic information, but this may be presented in a variety of ways, and their decoding in relation to the physical appearance of the object that they describe can prove challenging.

The following examples show different styles of collation formulas:

[1] i, 1-9 (8), 10 (6), 11-20 (8), 21 (7), i
[2] I-III^{8}, IV^{10}, V-IX^{8}

[3] IV(32), IV-1(40), 9 IV(120), IV-4
[4] $1\text{-}4^{8}$, 5^{2}, $6^{4\text{-}1}$, $7\text{-}10^{10}$
[5] 2°: $^{\pi}A^{6}$($^{\pi}$A1+1, $^{\pi}$A5+1.2), A-$2B^{6}$, $2C^{2}$, a-g^{6}, $^{\chi}2g^{8}$, h-v^{6}, x^{4}, “gg3.4”(±”gg3”), ¶-$2¶^{6}$, 3¶1, 2a-$2f^{6}$, $2g^{2}$, “Gg^{6}”, $2h^{6}$, 2k-$3b^{6}$

Of these, the first four illustrate different patterns of collation formulas utilized for manuscripts, whilst the latter shows a bibliographical description of the gathering assembly of a printed book.

Formulas to describe manuscripts and printed books aim at the same scope: representing the gathering structure of a book in codex format; there are, however, some fundamental differences between the two schools. In manuscript studies collation formulas represent book structures exactly as they are, whilst bibliographical formulas represent the ideal copy of the printed book, and not the state of specific exemplars. In addition, manuscript studies—unlike the case of printed books and their bibliographical description—lack a standard for drafting collation formulas that is approved and employed by all scholars. As it can be seen in the examples above—[1] to [4]—some schemas use Roman numerals to signal the sequence of gatherings, whilst others prefer Arabic numerals; some use superscripts, and some show the number of pages in a group. Without being familiar with specific schemas, the interpretation of manuscript collation formulas can be problematic. Nonetheless, for the most part, both bibliographical collation formulas and the various styles of those employed in manuscript studies share a set of information units that are necessary to describe the arrangement of the sheets within textblocks.

Zappella (1996) and Andrist et al. (2013) provide a comprehensive overview of the state of the art of collation formulas in bibliography and manuscript studies.

There have been some attempts to formulate collation schemas, and to model the gathering structure of books, in a way that such information could be easily parsed by computers. Gerardy (1972) describes a numerical system[1] to encode gathering structures of manuscripts. This collation format works like a decimal cataloguing system, and assigns numbers to gatherings (GG), bifolia (BB), and folios (f) or sides (s) of the leaves (i.e. recto and verso) according to a specific template:[2] GG.BB.f|s. This system assigns a unique numerical code to each element and accommodates for irregular structures by allocating special codes to stubs (7) and missing leaves (0). In this manner, also the frequent—and difficult to model—case of quires within quires, can be encoded. However, the unique numerical IDs in themselves do not communicate the relationships that exist amongst the bifolia within a gathering, i.e., looking at the example in figure 1, knowing that a folio ID is 03.03.01 and that of another is 03.04.01 does not convey the fact that bifolio 03 is an example of a quire

[1] See ‘Pagination décimale’ in Muzerelle 1985.
[2] The full template also accommodates for stubs and missing leaves, and not just full folios and their sides.

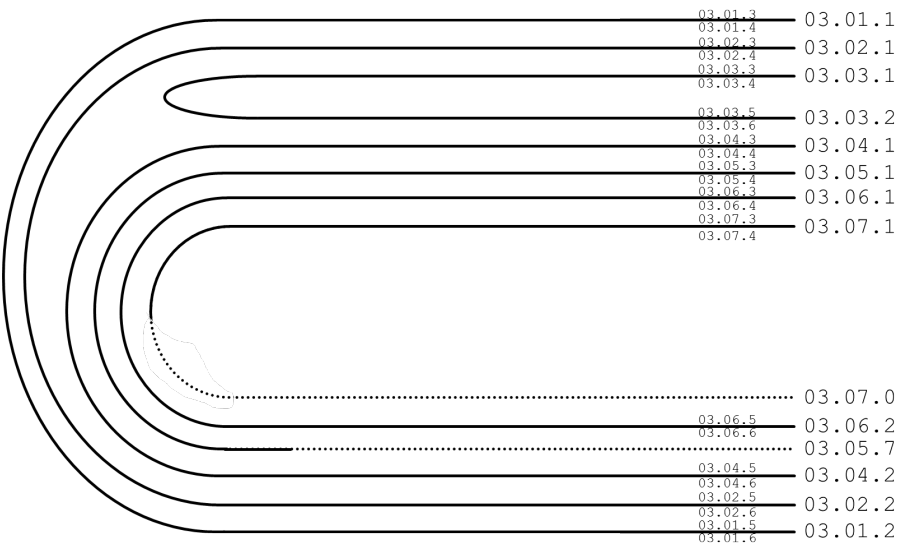


Figure 1: Example of decimal pagination for a complex gathering 3 in a manuscript (after Gruijs 1974, 254, schema 2; and Gerardy 1980, 45).

within quire: only the diagram or the full array of the gathering's IDs yield this important piece of information, and this is a significant flaw of the system.

In 2004, the TEI Physical Bibliography Workgroup put together a proposal to expand the collation recording capabilities of the TEI-MS model (TEI Workgroup on Physical Bibliography 2004). Considering that the physical structure of a book can be conceptualized as a series of hierarchically-organized objects, such as gatherings which contain leaves, and pages which contain lines of text, the working group advanced two distinct models to be integrated within a `<collation>` element in the `<msDescription>` or `<bookDescription>` of the TEI header.

On the one hand, in `<collationFormula>`, the typical layout of collation formula schemas was transposed within a hierarchical structure containing the elements that make up a full bibliographic description of gathering structures: a list of gatherings, an indication of the total number of leaves, pagination statements, etc.

On the other hand, the working group modelled a complex series of elements—i.e. `<gathering>`, `<leaf>`, `<page>`—to directly describe the physical structure of books in codex format.

This module did not become part of TEI P5 (TEI 2016b), and as a result, the standard way of recording collation information within TEI-based descriptions is to insert,

within a <collation> element, using informal prose, or other notational conventions, a description of a book's current and original arrangement of leaves and gatherings (TEI 2016a). The guidelines do not, therefore, prescribe any specific collation notation, but typical collation formulas can be included in a <formula> element as text. The ideas brought forward by the Physical Bibliography working group were, however, valuable, and, as it will be seen, they are being integrated in our own modelling of the gathering structures.

```
<gathering>
 <leaf xml:id="leaf1" conjunct="leaf8">
  <page xml:id="p1" sheetSide="1" cutFromN="p8" W="p16"/>
  <page xml:id="p2" sheetSide="2" cutFromN="p7" E="p15"/>
 </leaf>
 <leaf xml:id="leaf2" conjunct="leaf7">
  <page xml:id="p3" sheetSide="2" cutFromN="p6" W="p14"/>
  <page xml:id="p4" sheetSide="1" cutFromN="p5" E="p13"/>
 </leaf>
 <leaf xml:id="leaf3" conjunct="leaf6">
  <page xml:id="p5" sheetSide="1" cutFromN="p4" W="p12"/>
  <page xml:id="p6" sheetSide="2" cutFromN="p3" E="p11"/>
 </leaf>
 <leaf xml:id="leaf4" conjunct="leaf5">
  <page xml:id="p7" sheetSide="2" cutFromN="p2" W="p10"/>
  <page xml:id="p8" sheetSide="1" cutFromN="p1" E="p9"/>
 </leaf>
 <leaf xml:id="leaf5" conjunct="leaf4">
  <page xml:id="p9" sheetSide="1" cutFromN="p16" cutFromE="p12" W="p8"/>
  <page xml:id="p10" sheetSide="2" cutFromN="p15" cutFromW="p11" E="p7"/>
 </leaf>
 <!-- [...] -->
</gathering>
```

Listing 1: Example of encoding according to the 2004 Physical Bibliography proposal.

1.2 Viewing digitized manuscripts

Digitized medieval manuscripts are typically viewed through single-page or facing-page interfaces, which lack the physical cues present in a physical book, i.e., the size of the book, its thickness, details of the parchment or paper, etc. Indeed, even facing-page interfaces do not usually show a picture of book openings at all, but rather they are composites made with two images: one of the left-side page and another of the right-side page. These images would have been taken at different times. Typically all images of one side pages are taken first, e.g. all the rectos, then of the other side, and then file names or structural metadata are used to order the files correctly in post processing. Most digital libraries provide some information on the pages depicted, and views other than single-page or facing-page: all provide information on the folio number and the side (recto or verso) shown; some indicate the quire number (e.g. The British Library et al. 2016), and some offer a variety of viewing modes, including pages of thumbnails (e.g. *E-Codices* 2016) or thumbnails presented filmstrip-style

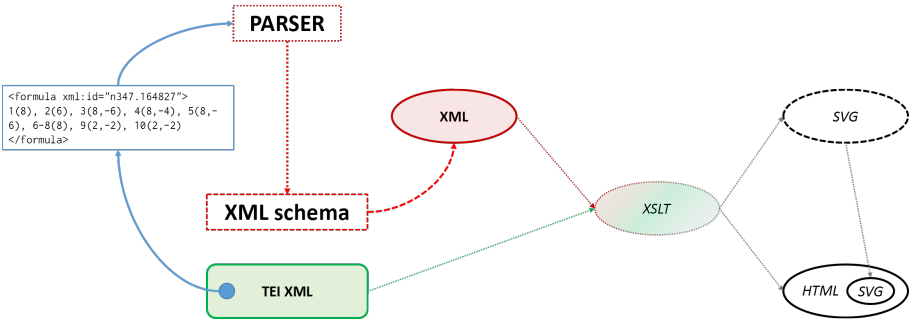


Figure 2: Diagram showing the pipeline of the prototype system.

across the bottom of a page (e.g. *Vitae Sanctorum* 2016). However, again, for the most part, the focus of these resources is on the page, rather than on the physical object. Even the Turning the Pages™ software (*Turning the Pages™* 2016), conceived by the British Library in 1996—and developed by Armadillo Systems (*Armadillo Systems* 2009) since 2001—, which, since version 2.0 (2006), has produced realistic three-dimensional books (including the ability to mimic the different movement of paper and parchment pages as these are turned), lacks any modelling of the gathering structure. To present knowledge, there is no institutional digital library that describes the physicality of manuscripts outside of the standard Physical Description section of the manuscript records and collation formulas.

In VisColl, we first model the collation of manuscripts in an XML format and then process that model in various ways, currently providing both diagrams and formulas, but potentially in other novel ways as well. For instance, in addition to visualizing the physical structure of a manuscript, the Beta Version of VisColl currently under development enables users to create taxonomies describing the content of the manuscript, and other elements, and then the system links those taxonomies to the physical structure, which produces a more robust and descriptive visualization than is possible in the current system.

This paper will document the stages of the development of VisColl, from its conception to its current instantiation, highlighting the steps taken and the reasoning behind each new actualization of the project. The current state of development can be found at the VisColl's GitHub page (Porter 2016a), which documents each new build, and from which the project's code can be downloaded.

2 Proof of concept

In July 2013, work started on the proof of concept for VisColl (cf. Porter 2013). This was established by taking an existing collation formula schema—i.e. that was devised

by William Noel (2011) for the Digital Walters project[3]—and processing it into two separate visualizations: quire diagrams (showing how leaves pair into bifolia) and what the project calls *Bifolia View*, where images of each page are viewed alongside the other half of the sheet as bifolia (useful in cases where it is not clear whether the sheets were written/illustrated before or after they were gathered into quires). In practice these two visualizations were presented together, with a quire diagram on the left side of a page and bifolia view presented to the right. Outside of digital practice, this perspective is only achieved by disbinding a manuscript. Figure 2 shows a diagram of the prototype pipeline: the collation formula (presented as text content of `<formula>`) was extracted from the TEI XML and parsed into XML. This collation XML was then processed along with the TEI XML, and the collation XML was converted into SVG diagrams while the image files, listed in the `<facsimile>` section of the TEI XML, were collected and arranged into bifolia. The bifolia are displayed on an HTML page with the SVG diagrams embedded alongside. Each quire was presented on its own web page using a combination of HTML for the page wrapper and SVG for the quire diagram. Each bifolia was presented in a row with the "active" bifolia highlighted in the diagram. The images were presented alongside the diagram: first, with the "inside" of the sheet facing up and then with the "outside" (as though the sheet were turned over; see fig. 3).

The great benefit of the proof of concept approach is that it enabled the batch processing of several manuscripts at once. At one point, Porter created visualizations in a single afternoon for all the manuscripts on the Digital Walters website that had associated collation formulas. There were, however, several downsides to this approach. The main problem is that it was entirely dependent on a specific collation formula schema. Unlike printed books, there is no single standard for manuscript collation formulas. Although all formulas will contain the same basic information, it may be presented in various ways.

3 Alpha version

The alpha version of VisColl had two main aims. The first, derived from the proof of concept, was to move from a formula-based approach to a model-based approach. The second was to build a system that would be publicly available and easily accessible. This second aim was a weakness of the proof of concept version. Although the proof of concept scripts were available on GitHub, with basic documentation on how to run them, it was difficult for users to run them correctly, especially as the scripts were not able to process any but the most basically constructed manuscripts. With this in mind, the alpha version system was built in two parts, with a third step that a user

[3] A website making available the digital images and metadata of the manuscripts held at the Walters Art Museum (*The Digital Walters* 2016).

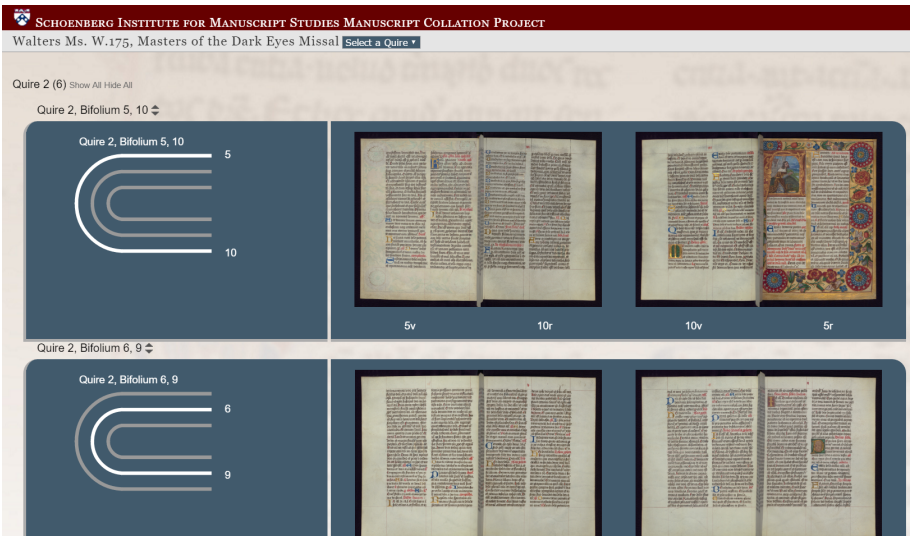


Figure 3: Screenshot of Baltimore, Walters Ms.175, which was visualized with the proof of concept prototype. Note the two views (inside/outside) for each bifolium, and the collation diagrams, highlighting which set of leaves are being visualized.

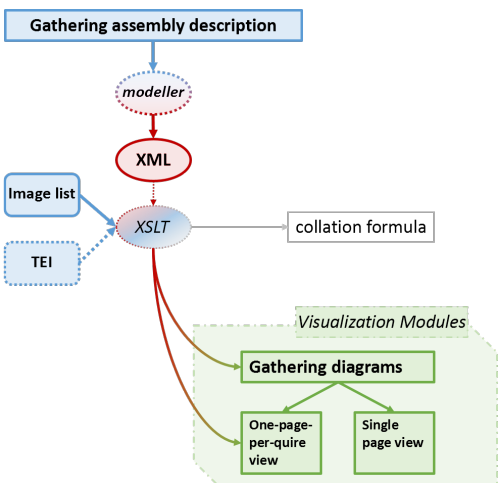


Figure 4: Diagram of the alpha version pipeline with its three steps (Collation Modeler, Image List, and Visualization Generation) and four XSLT outputs (collation formula, collation diagrams, one-page-per-quire visualization, single-page visualization).

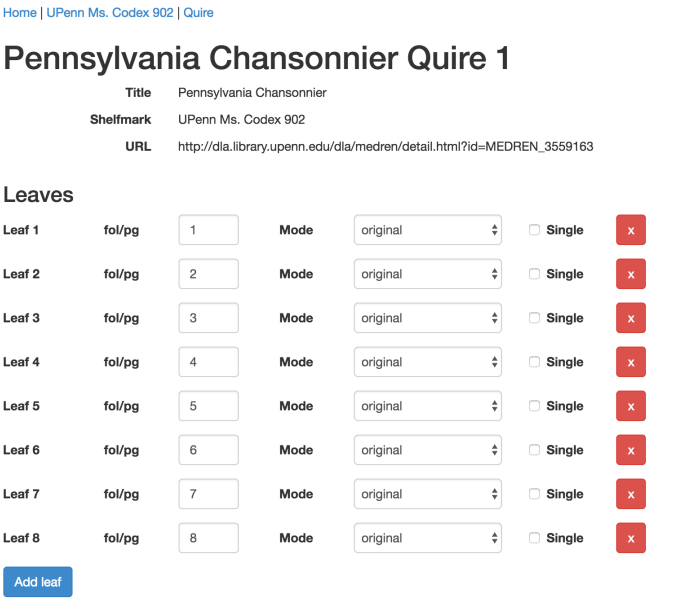


Figure 5: A screenshot of the Collation Modeler showing the complete construction of a quire for UPenn Ms. Codex 902.

would need to perform on their own. These three steps are: Collation Modeler, Image List, and Visualization Generation (see fig. 4).

3.1 Collation Modeler

The Collation Modeler,[4] built in Ruby on Rails by Doug Emery at the University of Pennsylvania, enables a user to construct and export a collation model, which is specifically formatted to be input into the Visualization Generation tool. In the current version of the Collation Modeler, using a form-based interface a user builds a number of quires and then identifies each leaf in the quire as original (to the manuscript), added (to the manuscript), missing (from the manuscript) or replaced (the original leaf having been removed and replaced with another leaf containing the same text as the original).

[4] The publicly accessible Collation Modeler (*Collation Modeler* 2016) and the Collation Modeler code on GitHub (Emery 2016).

An XML file containing the collation model is generated from the Collation Modeler to be used to create visualizations. In the current version, visualizations can't be generated directly from the Collation Modeler, which we recognize as a barrier for use.

```
<?xml version="1.0"?>
<manuscript>
 <url>http://dla.library.upenn.edu/dla/medren/detail.html?id=MEDREN_3559163</url>
 <title>Pennsylvania Chansonnier</title>
 <shelfmark>UPenn Ms. Codex 902</shelfmark>
 <quire n="1">
  <leaf n="1" mode="original" single="false" folio_number="1" conjoin="8"
       position="1" opposite="8"/>
  <leaf n="2" mode="original" single="false" folio_number="2" conjoin="7"
       position="2" opposite="7"/>
  <leaf n="3" mode="original" single="false" folio_number="3" conjoin="6"
       position="3" opposite="6"/>
  <leaf n="4" mode="original" single="false" folio_number="4" conjoin="5"
       position="4" opposite="5"/>
  <leaf n="5" mode="original" single="false" folio_number="5" conjoin="4"
       position="5" opposite="4"/>
  <leaf n="6" mode="original" single="false" folio_number="6" conjoin="3"
       position="6" opposite="3"/>
  <leaf n="7" mode="original" single="false" folio_number="7" conjoin="2"
       position="7" opposite="2"/>
  <leaf n="8" mode="original" single="false" folio_number="8" conjoin="1"
       position="8" opposite="1"/>
 </quire>
 <quire n="2">
  <leaf n="1" mode="original" single="false" folio_number="9" conjoin="8"
       position="1" opposite="8"/>
  <leaf n="2" mode="original" single="false" folio_number="10" conjoin="7"
       position="2" opposite="7"/>
  <leaf n="3" mode="original" single="false" folio_number="11" conjoin="6"
       position="3" opposite="6"/>
  <leaf n="4" mode="original" single="false" folio_number="12" conjoin="5"
       position="4" opposite="5"/>
  <leaf n="5" mode="original" single="false" folio_number="13" conjoin="4"
       position="5" opposite="4"/>
  <leaf n="6" mode="original" single="false" folio_number="14" conjoin="3"
       position="6" opposite="3"/>
  <leaf n="7" mode="original" single="false" folio_number="15" conjoin="2"
       position="7" opposite="2"/>
  <leaf n="8" mode="original" single="false" folio_number="16" conjoin="1"
       position="8" opposite="1"/>
 </quire>
 <quire n="3">
  <leaf n="1" mode="original" single="false" folio_number="17" conjoin="8"
       position="1" opposite="8"/>
  <leaf n="2" mode="original" single="false" folio_number="18" conjoin="7"
       position="2" opposite="7"/>
  <leaf n="3" mode="original" single="false" folio_number="19" conjoin="6"
       position="3" opposite="6"/>
  <leaf n="4" mode="original" single="false" folio_number="20" conjoin="5"
       position="4" opposite="5"/>
  <leaf n="5" mode="original" single="false" folio_number="21" conjoin="4"
       position="5" opposite="4"/>
  <leaf n="6" mode="original" single="false" folio_number="22" conjoin="3"
```

```
      position="6" opposite="3"/>
  <leaf n="7" mode="original" single="false" folio_number="23" conjoin="2"
      position="7" opposite="2"/>
  <leaf n="8" mode="original" single="false" folio_number="24" conjoin="1"
      position="8" opposite="1"/>
 </quire>
 <!-- [...] -->
</manuscript>
```

Listing 2: Example XML code of the collation model for UPenn Ms. Codex 902.

3.2 Image list

The image list is a file required by the Visualization Generation tool. If the user wants a bifolia view, the image list must include folio numbers or page numbers along with URLs to the corresponding image file. The system does not import these images, rather the HTML output points to the image files wherever they reside on the web. If the user does not need a bifolia view an image list file still needs to be uploaded to the Visualization Generation tool, but it may be an empty file.

In the alpha version, the image list needs to be built in an Excel spreadsheet with page/quire numbers in the first column and file URLs in the second column. The file is saved as an XML spreadsheet, and this file is fed into the Visualization Generation tool along with the collation model. The Beta Version will enable input in a TEI facsimile format, which would make it easier for someone working with TEI files.

3.3 Visualization Generation tool

The Visualization Generation tool is a web front-end built on top of an XSLT pipeline that uses XProc-Z, developed by Conal Tuohy (2016). The XSLT scripts are relatively

	A	B
1	1r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0006_web.jpg
2	1v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0007_web.jpg
3	2r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0008_web.jpg
4	2v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0009_web.jpg
5	3r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0010_web.jpg
6	3v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0011_web.jpg
7	4r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0012_web.jpg
8	4v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0013_web.jpg
9	5r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0014_web.jpg
10	5v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0015_web.jpg
11	6r	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0016_web.jpg
12	6v	http://openn.library.upenn.edu/Data/0001/ljs101/data/web/0241_0017_web.jpg

Figure 6: Example image list in Excel.

unchanged from the proof of concept version, except that the first few scripts (which parsed the collation formula into the proto-collation model) have been removed, as the processing now begins with the collation model exported from the Collation Modeler.[5] The final output script has also been changed, as it now outputs four different views instead of the single one-web-page-per quire view from the proof of concept. In addition to the one-page-per quire view, it is now possible to generate the following: a single page for the whole manuscript (quires can be viewed and hidden at will), a diagrams-only view without the bifolia view, and a collation formula. In order to use the Visualization Generation tool, the user must upload both a collation model and an image list. The system depends on the folio or page numbers in the image list and the collation model to match. In a few minutes, the system outputs a zip file containing all four visualizations.[6]

Even in its imperfect alpha version, VisColl is being used in the community of manuscript scholars. Most notably, Lisa Fagin Davis is using VisColl in her project to reconstruct the physical construction of the *Beauvais Missal*, a late thirteenth-century liturgical book that was dismembered in 1942, when individual leaves were sold to institutions and individuals throughout the USA. As of October 2016, Fagin Davis has successfully reconstructed four quires of this manuscript (Fagin Davis 2016). Furthermore, Dot Porter and Will Noel at the University of Pennsylvania have used VisColl in their class for the Rare Book School, "The Medieval Manuscript in the 21st Century", and their students have in some cases made new findings with assistance from the tool (McDowell 2015).

4 Beta version

We are currently working on the beta version of VisColl, with the collaboration of Alexandra Gillespie and her team at the Old Books, New Science (OBNS) Lab at the University of Toronto (Gilllespie and Mitchell 2016). In the beta version we will do three things. First, we will extend the model to include the definition of sets of terms (i.e., *taxonomies*) that users can use to describe both physical and textual aspects of manuscripts. Second, we will add a facility that enables users to link these terms to the physical components of the manuscript. Third, we are changing the physical model itself to be more flexible, and to enable more complex physical structures. The first two changes will be accomplished by creating two new sections in the model: a Taxonomies section, where vocabularies are defined and selected, and a Maps section,

[5] The alpha scripts are still available on GitHub (Porter 2015).

[6] Although the Visualization Generation tool does not allow for bulk processing, the scripts that run the tool are on GitHub and could be used to bulk process multiple collation models (Porter 2016b).

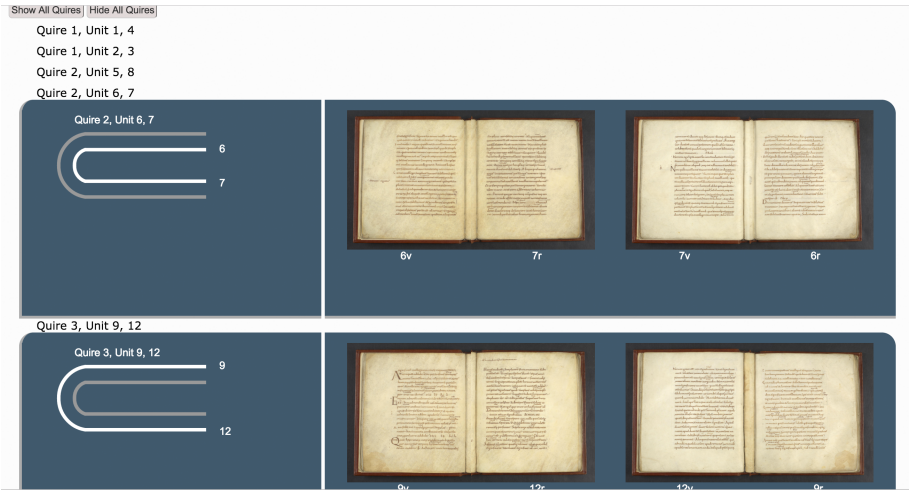


Figure 7: Screenshot of single-page view of University of Pennsylvania LJS 101, *Periermenias Aristotelis*. Note that all quires are on a single HTML page and the quires may be shown or hidden individually.

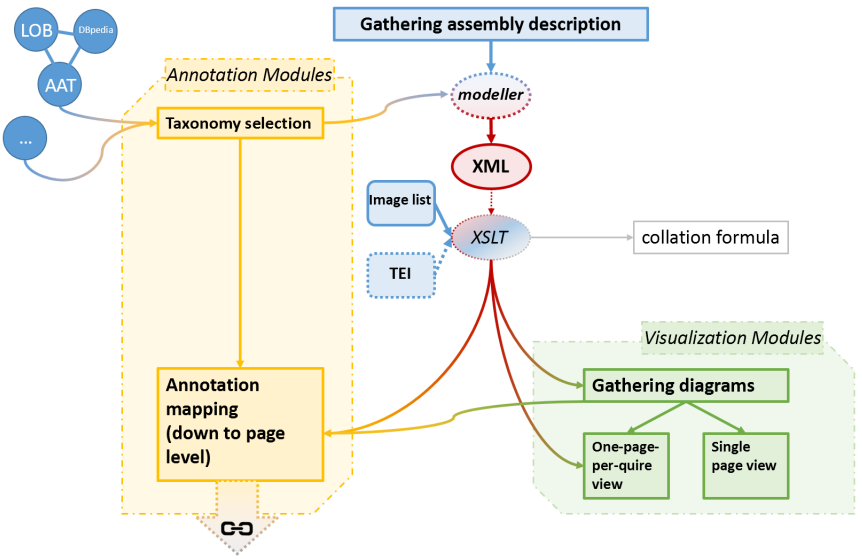


Figure 8: Diagram showing the pipeline of VisColl beta version. Note the integration of the Annotation modules and the possible links with external taxonomies and datasets.

where terms defined in the taxonomies section are linked to physical pieces of the manuscript.

4.1 Taxonomies section

In the Taxonomies section, users define lists of terms that describe important physical or textual aspects of the manuscript, and can then be associated with the physical components of the manuscript (sides of leaves, whole leaves, quires, and the entire manuscript). For example, if a manuscript is made of both parchment and paper, the user can define terms "paper" and "parchment", then in the model they can label each leaf with either term as appropriate. Taxonomies can include both defined by the project (e.g., the five stages of finish on the illustrations in the *Illustrated Hexateuch*, cf. Johnson 2000) and defined by external authorities (such as the Getty Art & Architecture Thesaurus, see The Getty Research Institute 2016; or the Language of Bindings Thesaurus, see Ligatus Research Centre 2016), opening the project to integration with Linked Data activities (Heath 2016). Any number of taxonomies can be defined in this section. Additionally, there are no taxonomies that are native to or required by the project. This is particularly important, as it allows for maximum flexibility on the side of the user, i.e., by selecting suitable taxonomy concepts, the user is able to describe anything in the model without restriction. In the example code below, the taxonomies section does not include values for semantic tags that are characteristic of the object, such as specific catchwords or signatures. However this is actively being addressed for inclusion in the final beta version of the model.

```
<viscoll>
 <taxonomy xml:id="b" xmlns="http://schoenberginstitute.org/schema/taxonomy">
  <!-- [...] -->
  <term xml:id="b5">Deuteronomy</term>
  <term xml:id="b6">Joshua</term>
 </taxonomy>
 <taxonomy xml:id="c">
  <label>Page contents</label>
  <term xml:id="c1">Illustration</term>
  <term xml:id="c2">Text</term>
 </taxonomy>
 <taxonomy xml:id="c"
           ref="http://www.getty.edu/research/tools/vocabularies/aat/">
  <label>Getty Art and Architecture Thesaurus</label>
  <term xml:id="c1" ref="http://vocab.getty.edu/aat/300011851">Parchment</term>
  <term xml:id="c2" ref="http://vocab.getty.edu/aat/300014179">Paper</term>
 </taxonomy>
 <taxonomy xml:id="d" ref="https://www.bl.uk/catalogues/illuminatedmanuscripts/
     glossary.asp">
  <label>Michelle P. Brown, Understanding Illuminated Manuscripts: A Guide to
       Technical Terms (J. Paul Getty Museum: Malibu and British Library: London,
        1994), online on the British Library website</label>
  <term xml:id="d1" ref="https://www.bl.uk/catalogues/illuminatedmanuscripts/
       GlossH.asp#hairside">Hair side</term>
```

```
  <term xml:id="d2" ref="https://www.bl.uk/catalogues/illuminatedmanuscripts/
       GlossF.asp#fleshside">Flesh side</term>
 </taxonomy>
 <taxonomy xml:id="e">
  <label>State of Finish (defined by Withers 2007)</label>
  <term xml:id="e1">Stage 1</term>
  <term xml:id="e2">Stage 2</term>
  <term xml:id="e3">Stage 3</term>
  <term xml:id="e4">Stage 4</term>
  <term xml:id="e5">Stage 5</term>
 </taxonomy>
 <!-- [...] -->
</viscoll>
```

Listing 3: Taxonomy section. Note that taxonomies are the responsibility of the user. They may be created by the user ("Page contents", "State of Finish") or may be drawn from formal schemas ("Getty Art & Architecture Thesaurus", "Understanding Illuminated Manuscripts").

4.2 Mapping section

The Mapping section links terms defined in the Taxonomy section to the physical components of the manuscript: sides of leaves, whole leaves, quires, or the entire manuscript. This creates reference links between semantic tags and physical components of the manuscripts. In the working version (see listing 4) the map identifies leaves by quire number and leaf in the quire (i.e., the third leaf of quire one is identified as `@leaf="1.3"`) and the side is indicated by `@side="r"` or `@side="v"`. Moving forward we will replace this physical identification with pointers to unique identifiers in the collation model, and thus the map will simply be a space for linking together physical components and terms, rather than defining the physical components in any way itself.

```
<mapping>
 <map leaf="1.2" side="r">
  <term target="#c2 #b1 #e5"/>
 </map>
 <map leaf="1.2" side="v">
  <term target="#c2 #e5 #b1"/>
 </map>
 <map leaf="1.3" side="r">
  <term target="#c1 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.3" side="v">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.4" side="r">
  <term target="#c1 #c2 #e5 #b1 #d2"/>
 </map>
 <map leaf="1.4" side="v">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.5" side="r">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
```

```
 </map>
 <map leaf="1.5" side="v">
  <term target="#c1 #c2 #e5 #b1 #d1"/>
 </map>
 <map leaf="1.6" side="r">
  <term target="#c1 #c2 #e5 #b1 #d2"/>
 </map>
 <map leaf="1.6" side="v">
  <term target="#c1 #c2 #e5 #b1 #d3"/>
 </map>
 <!-- [...] -->
</mapping>
```

Listing 4: Mapping section links the taxonomies (the values of `@target`) to quires, leaves, and pages. The next version of the collation map will assign unique identifiers to leaves and these ids will be used in the map.

The taxonomy and mapping modules allow for the expansion of VisColl beyond the presentation of information, and permit the end user to add knowledge in a way that is directly linked with the physicality of manuscripts. This kind of annotation on a page-by-page basis is not novel per se in manuscript studies (cf. Németh 2015, 309-12, table 6, and Corbach 2013, 27-33, table 1), but for the first time, with VisColl, annotations can be added electronically and then consistently linked with the appropriate parts of manuscripts. Additionally, allowing the use of externally defined Link Data taxonomies fosters collaboration and opens data for further research beyond specific manuscripts and repositories, breaking data free of information silos.

4.3 Collation model

Additionally, at the time of writing, the XML schema behind the Collation Modeler is being totally restructured. Moving away from quires as basic units, the new model considers leaves and stubs as atomic elements—which together form folios, bifolia, quires, and bookblocks—in order to accommodate for those complex structures, a sign of the complicated lives, often found in manuscripts; structures such as that depicted in figure 9, with quires within quires and pasted singletons, are rarely (if ever) encoded in collation formulas. Another element that is not encoded in formulaic quire assembly descriptions, but that is indispensable to understand non-standard and complex quire structures, is the leaf attachment method. Leaves can either be sewn or pasted/glued together to form quires and bookblocks. The new model provides means to indicate the attachment method of each leaf (sewn being the default option), and this will in turn allow the end user to describe and visualize unambiguously exceptionally complex structures.

Finally, in the near future, it is hoped that the collation model within VisColl and its visualization and annotation modules might be integrated with the International Image Interoperability Framework (IIIF - 2016), since such a partnership would be

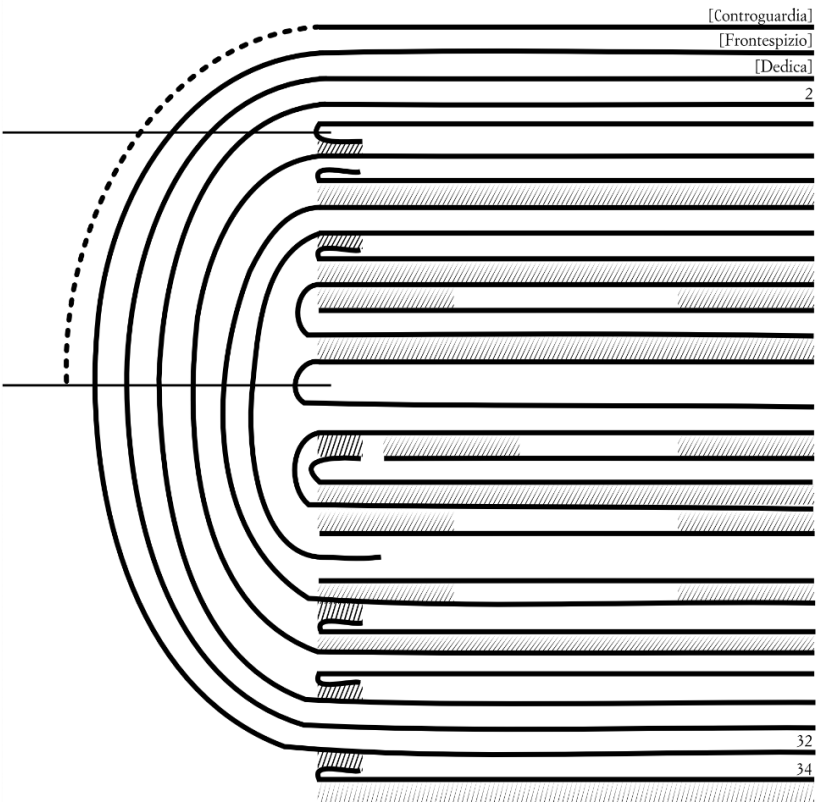


Figure 9: Example of complex manuscript quire structure (Vatican Library, Ferr.208, quire 1). Sewn leaves are indicated by a line representing the sewing thread; shaded areas indicate pasted leaves.

beneficial to both projects. Currently, the IIIF presentation API (Appleby et al. 2012) leverages the Shared Canvas Model (Sanderson and Albritton 2013) and the Web Annotation Data Model (Sanderson et al. 2017), and this accommodates annotation practices, which, by virtue of integrating seamlessly the principles of Linked Data and the Architecture of the Web (Jacobs and Walsh 2004), are perhaps more robust than VisColl annotations alone. The Shared Canvas Model, however, is incapable of representing the connections between different canvases (i.e. different pages of a manuscript) beyond being in a sequence. Integrating this representation model with the VisColl collation model would preserve both IIIF's robust annotation procedures and VisColl's effective representation of the actual structure of codices.

5 Conclusions

Since its initial conception in the mid-2000s through its implementation in 2013 up to current work on the beta version, VisColl, with its conceptual design and front-end usability, has been developed primarily for scholars who work with manuscripts. The project has brought together manuscript scholars, librarians and curators, conservators, and software developers, and serves as an example of the synergistic outcomes possible with interdisciplinary collaboration. Collaboration has increasingly brought flexibility into the project, widening its scopes to accommodate a diverse range of activities typical of specific disciplines that have the study of manuscripts at their core. This should not come as a surprise since the quire assembly is central to the production of codices, and its study is therefore fundamental for all disciplines within manuscript studies (and beyond). This collaborative effort will continue as we finalize the back-end design and modeling challenges and through the development of more ways to effectively visualize the new data brought into the beta version.

Bibliography

Andrist, Patrick, Paul Canart, and Marilena Maniaci. *La syntaxe du codex: essai de codicologie structurale.* Bibliologia 34. Turnhout: Brepols, 2013.

Appleby, Michael et al. (eds.) *IIIF Presentation API 2.1.* IIIF Consortium. 2012. <http://iiif.io/api/presentation/2.1>.

Armadillo Systems. London: Armadillo New Media Communications Ltd. 2009. <http://www.armadillosystems.com>.

Collation Modeler. 2016. <https://protected-island-3361.herokuapp.com>.

Corbach, Almuth. "Der Bernward-Psalter im Wandel der Zeiten. Eine Studie zu Ausstattung und Funktion." In Müller, Monika E. (ed.). *Der Bernward-Psalter im Wandel der Zeiten: Eine Studie zu Ausstattung und Funktion.* (=Wolfenbütteler Mittelalter-Studien 23). Wiesbaden: Harrassowitz, 2013. 263–382.

E-Codices - Virtual Manuscript Library of Switzerland. Fribourg: University of Fribourg, 2016. <http://www.e-codices.unifr.ch/en>.

Emery, Doug. *Collation Modeling.* GitHub 2016. <https://github.com/demery/collation-modeling>.

Fagin Davis, Lisa. "Quire Visualizations." In *Reconstructing the Beauvais Missal.* Cambridge (MA): The Medieval Academy, 2016. <https://brokenbooks2.omeka.net/exhibits/show/quire-visualizations>.

Ferrajoli 208. Rome: Vatican Library.

Gerardy, Theo. "Die Beschreibung der Wasserzeichen in Manuskripten und Drucken." In International Association of Paper Historians (eds.). *XIe Congrès International, Arnhem (Hollande) 4–9 Juin 1972.* Haarlem: Stitching Papiergeschiedenis, 1972. 1–9.

Gerardy, Theo. "Die Beschreibung des in Manuskripten und Drucken vorkommenden Papiers." In Gruys, Albert and Johan Peter Gumbert (eds.). *Les Matériaux Du Livre Manuscrit.* Codicologica 5. Leiden: Brill, 1980. 37–51.

Gilllespie, Alexandra and Laura Mitchell. *Old Books New Science (OBNS) Lab.* Toronto: Centre for Medieval Studies, 2016. <https://oldbooksnewscience.com>.

Gruijs, Albert. "Le Protocole de Restauration et La Description Des Cahiers et Bifolia." In Glénisson, Jean and Louis Hay (eds.). *Les Techniques de Laboratoire Dans L'ètude Des Manuscrits: [Actes Du Colloque International] Paris, 13–15 Septembre 1972.* Colloques Internationaux Du Centre National de La Recherche Scientifique 548. Paris: Centre national de la recherche scientifique, 1974. 253–255.

Heath, Tom. *Linked Data - Connect Distributed Data across the Web.* Linked Data community, 2016. <http://linkeddata.org>.

International Image Interoperability Framework. IIIF Consortium. 2016. <http://iiif.io>.

Jacobs, Ian and Norman Walsh (eds.). *Architecture of the World Wide Web, Volume One.* W3C, 2004. <https://www.w3.org/TR/webarch>.

Johnson, David. "A Program of Illumination in the Old English Illustrated Hexateuch: *Visual Typology.*" In Barnhouse Rebecca, and Benjamin C. Withers (eds.). *The Old English Hexateuch: aspects and approaches.* Kalamazoo (MI): Medieval Institute Publications, Western Michigan University, 2000. 165–200.

Kiernan, Kevin S. *Beowulf and the Beowulf Manuscript.* New Brunswick (NJ): Rutgers University Press, 1981.

Ligatus Research Centre. *Language of Bindings.* London: University of the Arts London, 2016. <http://www.ligatus.org.uk/lob>.

McDowell, Jesse. *An Ideal Collation of LJS 101.* November 16 2015. <http://schoenberginstitute.org/2015/11/16/an-ideal-collation-of-ljs-101>.

Muzerelle, Denis. *Vocabulaire codicologique: répertoire méthodique des termes français relatifs aux manuscrits.* (=Rubricae: Histoire du livre et des textes 1). Paris: Éditions CEMI, 1985.

Németh, András. "Layers of Restorations: Vat. Gr. 73 Transformed in the Tenth, Fourteenth, and Nineteenth Centuries". In *Miscellanea Bibliothecae Apostolicae Vaticanae* XXI, 2015. 281–330.

Noel, William. "Collation." In *The Digital Walters: Describing Manuscripts with TEI.* Baltimore (MD): Walters Art Museum, 2011. <http://thedigitalwalters.org/Supplemental/

ManuscriptDescription.html#collation>.
Porter, Dot. [2013.] "Visualizations of TEI Ms Descriptions." *tei-l@listserv.brown.edu*. 2013. <https://listserv.brown.edu/archives/cgi-bin/wa?A2=tei-l;775d4091.1307>.
— [2015.] "XSLT Alpha." In *Visualizing Physical Manuscript Collation*. GitHub. 2015. <https://github.com/leoba/VisColl/tree/master/xsl/xslts-alpha>.
— [2016a.] *Visualizing Physical Manuscript Collation*. GitHub. 2016. <https://github.com/leoba/VisColl>.
— [2016b.] "XSLTs." In *Visualizing Physical Manuscript Collation*. GitHub. 2016. <https://github.com/leoba/VisColl/tree/master/xsl/xslts>.
Sanderson, Robert, and Benjamin Albritton (eds.). *Shared Canvas Data Model 1.0*. IIIF Consortium. 2013 <http://iiif.io/model/shared-canvas/1.0>.
Sanderson, Robert, Paolo Ciccarese, and Benjamin Young (eds.). *Web Annotation Data Model*. W3C. 2017. <https://www.w3.org/TR/annotation-model>.
TEI. [2016a.] "10.7.1 Object Description." In *Text Encoding Initiative P5 (v. 3.0.0): Guidelines for Electronic Text Encoding and Interchange*, P5 revised and re-edited edition. Oxford, Providence (RI), Charlottesville (VA), Nancy (KY): Text Encoding Initiative Consortium, 2016. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html#msph1>.
— [2016b.] *Text Encoding Initiative P5 (v. 3.1.0): Guidelines for Electronic Text Encoding and Interchange*. P5 revised and re-Edited edition. Oxford, Providence (RI), Charlottesville (VA), Nancy (KY): Text Encoding Initiative Consortium, 2016. <http://www.tei-c.org/Guidelines/P5/>.
TEI Workgroup on Physical Bibliography. *Physical Bibliography - Draft for P5*. 2004. <http://www.tei-c.org/Activities/Workgroups/PB/PB-draft.xml>.
The British Library, National Library of Russia, St. Catherine's Monastery, and Leipzig University Library. *Codex Sinaiticus: See the Manuscript*. London: The British Library, 2016. <http://www.codexsinaiticus.org/en/manuscript.aspx>.
The Digital Walters. Baltimore (MD): Walters Art Museum, 2016. <http://thedigitalwalters.org>.
The Getty Research Institute. *Art & Architecture Thesaurus® Online*. Los Angeles (CA): The J. Paul Getty Trust, 2016. <http://www.getty.edu/research/tools/vocabularies/aat>.
Tuohy, Conal. *XPro-Z. A Platform for Running XProc Pipelines as Web Applications in a Java Servlet Container*. GitHub. 2016. <https://github.com/Conal-Tuohy/XProc-Z>.
Turning the Pages™. London: Armadillo Systems, 2016. <http://ttp.onlineculture.co.uk>.
"Vitae Sanctorum." In *Beinecke Digital Collections*. New Haven (CT): Yale University Library, 2016. <http://brbl-dl.library.yale.edu/vufind/Record/3592236>.
Withers, Benjamin C. *The Illustrated Old English Hexateuch, Cotton Claudius B.iv: The Frontier of Seeing and Reading in Anglo-Saxon England*. London, Toronto, Buffalo: The British Library ; University of Toronto Press, 2007.
Zappella, Giuseppina. *Manuale del libro antico*. Milano: Editrice Bibliografica, 1996.

Digitale Paläographie

Digital Palaeography

Advances in Handwritten Keyword Indexing and Search Technologies

Enrique Vidal

Abstract

Many extensive manuscript collections are available in archives and libraries all over the world, but their textual contents remain practically inaccessible, buried under thousands of terabytes worth of high-resolution images. If perfect or sufficiently accurate text-image transcripts were available, textual content could be indexed directly for plaintext access using conventional information retrieval systems. But the results of fully automated transcriptions generally lack the level of accuracy needed for reliable text indexing and search purposes. Additionally, manual or even computer-assited transcription is entierely unsustainable when dealing with the extensive image collections typically considered for indexing. This paper explains how accurate indexing and search commands can be implemented directly on the digital images themselves without the need to explicitly resort to image transcripts. Results obtained using the proposed techniques on several relevant historical data sets are presented, clearly supporting the considerable potential of these technologies.

Zusammenfassung

Auf der ganzen Welt halten Archive und Bibliotheken umfangreiche Sammlungen handschriftlicher Dokumente bereit. Doch bleiben deren Inhalte praktisch unzugänglich, verborgen unter tausenden von Terabytes hochaufgelöster Bilder. Gäbe es gute oder halbwegs verlässliche Text-Bild-Trankriptionen, ließen sich die jeweiligen Inhalte über herkömmliche Systeme zur Informationsrückgewinnung direkt indizieren und somit Zugänge zu entsprechenden Plaintext-Fassungen ermöglichen. Leider sind die Ergebnisse voll-automatisierter Transkriptionsverfahren zu ungenau, als dass sie sich für eine zuverlässige Textindizierung und Suche eigneten. Hinzu kommt, dass manuelle oder gar computergestützte Transkriptionsverfahren keine Nachhaltigkeit aufweisen, gerade wenn es sich um Bildsammlungen handelt, die aufgrund ihres großen Umfangs für eine Indizierung in Betracht gezogen werden. Dieser Artikel erläutert, wie verlässliche Indizierungen und Suchfunktionen unmittelbar auf den Bilddigitalisaten implementiert werden können, ohne dass dafür auf Bildtranskriptionen zurückgegriffen werden muss. Es werden Ergebnisse vorgestellt, die unter Anwendung der hier vorgestellten Technologie auf verschiedene historisch

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 103–119.

bedeutsame Datensätze erzielt worden sind und deren erhebliches Potential klar unter Beweis stellen.

1 Introduction

Handwriting is, in a way, like speaking, but in contrast to spoken language the written word has the property that it does not vanish immediately as it is preserved in textual form. In the centuries since humanity discovered such a convenient way of persistent communication, large amounts of handwritten documents have been produced. In fact, it is argued that the accumulated amount of handwritten text so far is larger than the available amount of machine-written text today (copies excluded), including modern digital-born text. Notwithstanding the questionability of this conjecture, it is fairly probable that our current knowledge of the history of human societies, based on the infinitesimal amount of handwritten text that has painfully been transcribed, might be rather limited.

In recent years, large quantities of historical manuscripts have been digitised and made available through web sites of libraries and archives all over the world. As a result of these efforts, many massive *image* collections of textual documents are available online. Irrespective of these efforts and the interest in their products, unfortunately these digitisations are largely useless for their primary purpose: exploiting the wealth of information conveyed by the text captured in the images. Therefore, there is a fast growing interest in automated methods which allow users to search for relevant textual information contained in these images.

In order to use classical text information retrieval approaches, a first step would be to convert the text images into digital text. Then, image's textual content could directly be indexed for plaintext access. However, OCR technology is completely useless for typical handwritten text images - and the results of fully automated transcriptions obtained using state-of-the-art *handwritten text recognition* (HTR) techniques lack the level of accuracy needed for reliable text indexing and search purposes (Vinciarelli et al. 2004; Graves et al. 2009; Romero et al. 2012).

An alternative to fully automatic processing is to rely on *computer-assisted* transcription. This was successfully explored empirically by Toselli et al. (2017), Romero et al.(2012) and Alabau et al. (2014), following new, powerful concepts of pattern recognition-based human-machine interaction introduced by Vidal et al. (2007) and Toselli et al. (2011). Following the positive results of these laboratory studies, preliminary evaluation by real users was carried out by Toselli et al. (2016). In this case, a historical botany book of about one thousand pages was fully transcribed interactively in less than three months by a team composed of one paleographer and four paleography students. In the past four years, the tranScriptorium (Transcriptorium)

project, has further explored the capabilities of these automatic and interactive HTR (IHTR) technologies to accelerate the conversion of raw text images into electronic text. These successful studies are now being continued within the recently started READ project.

Working conclusions from all studies mentioned above state:

a) To some extent, fully automatic transcripts of text images can be useful for plaintext indexing and search purposes. However, in many historical text image collections of interest, the typical level of transcription accuracy achieved severly hinders the search *recall*; i.e., the system's ability to ensure that all or most of the images which contain a given query text can actually be retrived is limited.
b) Similarly, the fully automatic transcription of most historical text images does not reach the level of accuracy needed for typical scholarly editions of the corresponding image collections.
c) In both cases, the required level of accuracy can obviously be obtained by means of additional user effort. If manual editing work is to be done, rather than just letting the users edit the noisy automatic transcripts, IHTR can be used to cost-effectively provide the desired transcription accuracy.
d) IHTR can significant reduce manual efforts regarding the edition of the automatic transcripts. But the overall effort demanded by IHTR is still substantial. Therefore, while IHTR is proving useful to produce scholarly editions of moderately sized historical collections, the required effort to handle extensive image collections targeted by indexing and search commands is entierely unsustainable.

This situation raises the need of search approaches specifically designed for large text *image* collections. In these approaches, on the one hand, indexing and search must be directly implemented in the images themselves, without explicitly resorting to image transcripts. On the other hand, rather than "exact" searching (as possible in plaintext), search queries have to be performed with a *confidence threshold*, somehow specified by the user as part of the query in order to meet the *precision-recall trade-off* which is considered most adequate in each query.[1]

Clearly, such a confidence-based query model cannot be properly implemented just by using conventional textual information retrieval methods on the noisy output of an automatic HTR system. Therefore, recognition techniques are needed which attach confidence measures to alternative word recognition hypotheses. Keyword spotting (KWS)[2] is a traditional way to address search problems within this framework. More

[1] Depending on the application, confidence thresholds can be specified more or less explicitly. For instance, in cases where results are provided in the form of ranked lists, the threshold is indirectly defined by the size of the list.

[2] See Manmatha et al. 1996; Rath and Manmatha 2007; Cao et al. 2009; Rodríguez-Serrano and Perronnin 2009; Kamel 2010; Fischer et al. 2012; Frinken et al. 2012; Wshah et al. 2012; Toselli and Vidal 2013a;

precisely, KWS aims to determine locations on a text image collection which are likely to contain an instance of a queried word, without explicitly transcribing the images.

KWS is generally qualified as a Query-by-Example (QbE) or a Query-by-String (QbS), depending on whether the query word is specified by means of an example-image or just as a character string respectively. While the QbE scenario can be useful in some applications, it is clearly not adequate for our purposes of indexing and search in large image collections. Therefore, in this paper we adopt the QbS framework. Moreover, it has been shown by Vidal et al. (2015) that highly accurate QbS performance can be achieved easily by exclusively using QbS technology.

Traditional work on handwritten KWS assumed previous segmentation of the text images into word image regions. However, word pre-segmentation is impossible for millions of historical manuscript images of interest and, even in favorable cases, it is quite prone to errors (Manmatha and Rothfeder 2005; Papavassiliou et al. 2010) which generally result in poor overall KWS performance (Ball et al. 2006). To overcome this important drawback, recent works[3] assume the (non-segmented) *line image* as the lowest search level. This is a convenient setting because, in most cases, text images can be segmented fully automatically into lines with appropriate accuracy (Papavassiliou et al. 2010; Bosch et al. 2012) and lines are sufficiently precise as target image positions for most practical textual image search and retrieval applications. Nevertheless, a fixed line segmentation can also be problematic in many cases and is nowadays considered perhaps the most severe bottleneck to achieve fully automatic processing of handwritten images for KWS and HTR alike. For this reason, our current work aims at indexing full pages in an attempt to circumvent the need for any kind of image segmentation alltogether.

On the other hand, most of the techniques which have been proposed for KWS can be considered to belong to one of these two broad classes: *training-based* and *training-free*. Training-based KWS methods are generally based on statistical optical (and language) models and typically adopt the QbS paradigm. Conversely, most training-free techniques are based on direct (image) template matching and assume the QbE framework.

The approaches proposed here are training-based and therefore need a certain amount (tens to hundreds) of manually transcribed images to train the required optical and language models. Additionally, they may benefit from the availability of collection-dependent lexicons and/or other specific linguistic resources. Our target applications are those involving large handwritten collections, where the effort or cost to produce these resources would pay off the benefits of making the textual contents of these collections readily available for exploration and retrieval.

Puigcerver et al. 2016; Toselli et al. 2016.

[3] See Kolcz et al. 2000; Terasawa and Tanaka 2009; Fischer et al. 2012; Frinken et al. 2012; Wshah et al. 2012; Toselli and Vidal 2013a; Toselli et al. 2016.

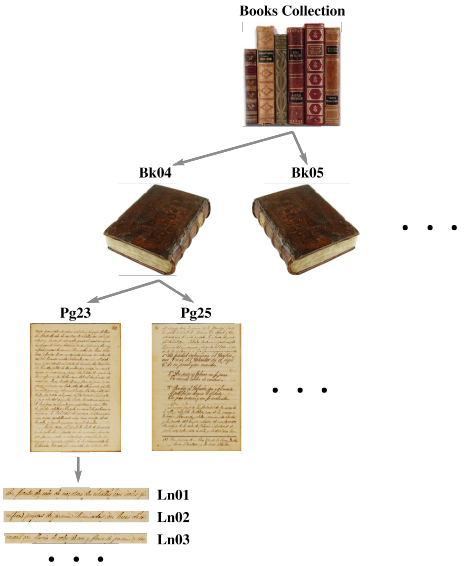


Figure 1: A hierarchical indexing and search model for handwritten text image collections. The top level in this illustration is a collection of books and the lowest level are line-shaped image regions. The specific levels of a hierarchy should be defined according to the characteristics of the document collection and search task considered.

2 Proposed indexing and search technology

An overview of the ideas behind the indexing and search technology we are developing is presented in this section. As previously stated, this technology assumes the *precision-recall trade-off search model* which requires *word confidence measures* computed for adequate regions of the text images of interest. Firstly, I will elaborate how these regions can be conveniently organized hierarchically and later I will explain how the required word confidence measures are computed.

A hierarchical indexing model

Indexing extensive document collections clearly calls for a hierarchical organization of indices. The lowest hierarchical level should consist of sufficiently small and meaningful *image regions*, such as text blocks (paragraphs) or lines. Figure 1 illustrates these concepts.

This kind of hierarchical organization of searchable text image regions entails important demands for the underlying precision-recall trade-off search model. Specifically, the word confidence measures must be defined properly, not only at the lowest level of the image region, but at every level of the hierarchy. In addition, con-

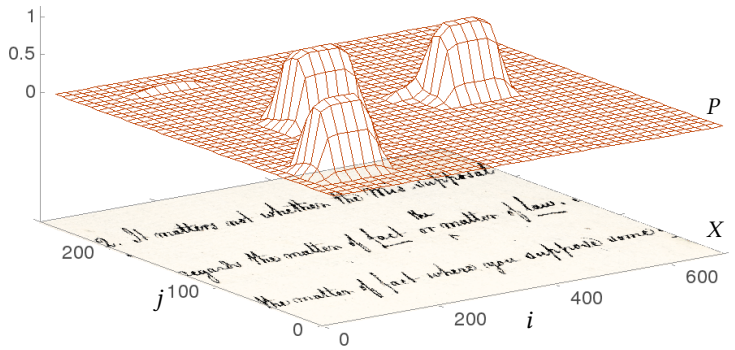


Figure 2: Pixel-level posteriorgram, P, for a text image X and word v ="matter".

fidence measures must be properly normalized and homogeneous across hierarchy levels. Clearly, when a user is searching for a certain word his or her intuition about what a confidence level of e.g. 0.7 (70%) means should be the same whether he or she is searching for books in a book collection, for pages of a book, or for specific lines on one page of this book. This stands in direct opposition to the much less demanding confidence measure requirements entailed by conventional *flat indexing models*, which typically aim only to produce a ranked list of probable image regions retrived for each given query.

To fulfil the requirements discussed above, our techinques are being developed within a sound *statistical KWS framework* which supports the computation of confidence measures with the required properties, as explained below.

Pixel-level word confidence measures: the "posteriorgram"

The proposed approach relies on the basic concept of a *pixel-level "posteriorgram"*. In a nutshell, this is a probability map computed for a given image X and a possible query word v. At each position (i, j) of X, the posteriorgram provides the posterior probability that the word v is written in some subimage of X which includes the pixel (i, j). Figure 2 illustrates this concept.

The value of P at each image position (i, j) can be easily obtained by statistical *marginalization*. Simply put, the idea is to consider that v may have been written in any possible bounding box of the image X which includes the pixel (i, j). The marginalization process simply adds the word recognition probabilities for all these bounding boxes. This means that a posteriorgram could simply be obtained by repeated application of any word classification system capable of recognizing isolated (pre-segmented) words. It goes without saying, however, that the better the classifier, the better the corresponding posteriorgram estimates.

Directly obtaining a full pixel-level posteriorgram in this way entails a formidable amount of computation. However, as will be discussed later, it can be efficiently computed by clever combinations of subsampling of the image positions (i, j) and adequate choices of the marginalization bounding boxes.

In our approaches we use fully fledged holistic HTR systems to compute the required isoloated word probabilities. This allows us to take advantage of the linguistic context to obtain accurate word classification probabilities. In figure 2, a contextual word classifier based on an n-gram language model was used to compute P for the word `"matter"`. This query led to comparatively low probabilities of X around ($i = 100$, $j = 200$) in an region in which the similar (but different) word `"matters"` appears. According to the language model, the 2-grams `"the matter"` and `"matter of"` are highly predictable, thereby boosting the probability that the word `"matter"` exists in these exact image regions. Conversely, the 2-grams `"It matter"` and `"matter not"` are highly unlikely, resulting in low pixel probabilities in the image region where the different word `"matters"` appears (the results would roughly be reversed should the query word be `"matters"` instead).

Image region word confidence measures

Posteriorgrams can be used directly for KWS: given a confidence threshold τ, a word v is only spotted in image positions (i, j) where P is bigger than τ. Altering this threshold, adequate *precision–recall* trade-offs can be achieved. However, this approach is not feasible for large image collections as indexing word confidences for every image pixel would be impossible. For indexing purposes, what we really need is the confidence that a word v is written within a pre-specified image region such as a line, a column, or a full page, without explicitly taking into account the exact location of a word in this specific region or the number of locations in which the word may appear. In information retrieval terminology, this is called *"relevance"*. For each image region to be indexed we need to obtain the probability that it is *relevant* for the given query word.

The process of exactly computing relevance probabilities can become a complex endeavor. Nevertheless, a comparatively simple and intuitively appealing approach is to compute the region relevance probability for a word v just as the maximum pixel-level probability for v over the whole region. For instance, if the whole X in figure 2 were considered a region to be indexed, the probability that X is relevant for the query `"matter"` is adequately approximated by the maximum of the four picks of the posteriorgram illustrated in this figure.

Choosing adequate minimal searchable image regions: line-level KWS

In our work so far, line-shaped regions have been adopted as the smallest and hierarchically lowest image elements to be indexed. From the user perspective, lines are target image positions sufficiently precise for most document image search and retrieval applications. From a technical perspective, on the other hand, line-shaped image

regions are particularly useful as they enable efficient computation of posteriorgrams by adequately choosing the bounding boxes needed for the underlying marginalization process. Moreover, in many cases text lines are fairly regular and standard line segmentation techniques can be used to automatically determine line-shaped image regions with fair accuracy. Finally, and most importantly, line-shaped text image regions typically contain most[4] of the relevant lingusitic context needed for precise computation of word classification probabilities using a recognizer based on a language model, as discussed below.

Efficient computation of posteriorgrams and relevance probabilities
In our approach, line-level posteriorgrams are computed most efficiently using *Word Graphs* which are generated as a byproduct of recognizing full line region images with a fully-fledged holistic HTR system based on *optical character models* and (N-gram) *Language Models* (Toselli et al. 2016). When applied to a line-shaped image region, these systems can take full advantage of the lingusitic context to provide accurate word classification probabilities. On the other hand, a WG obtained in this manner provides a large number of alternative horizontal word-level segmentations. These segments directly define adequate sets of bounding boxes; just as those required by the marginalization process used to compute the posteriorgrams.

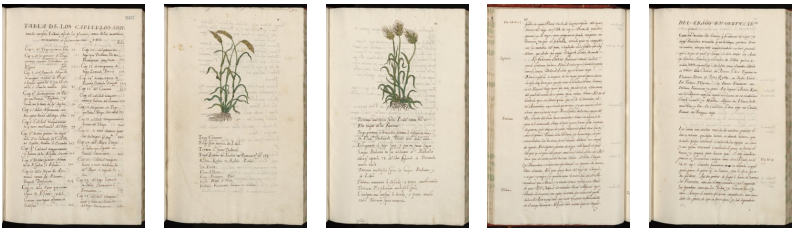
Line-region relevance probabilities are directly computed from the corresponding posteriorgrams, as explained above. They can in turn be combined easily and consistently to obtain *page-level* relevance probabilities (such as ... for *chapters*, *books*, etc., as needed for *hierarchical indexing*).

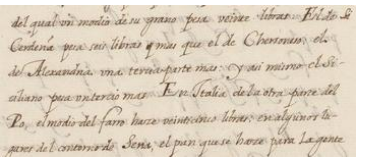
3 Laboratory results

Many collections of historical manuscript images have been considered for testing the proposed indexing and search technologies. Most of our research was pursued within the tranScriptorim project mentioned in section 1. The features of the data sets used in the experiments, the assessment measures adopted, and the results obtained, are presented in this section.

Data sets
A description summary and examples of the different data sets used in the experiments mentioned in this paper is given in figures 3–7. The first three data sets (Plantas, Bentham and Austen) correspond to collections which are comparatively modern (XVII-XIX century), entailing similar, relatively minor challenges in terms of writing style, homogeneity and language use. The last two data sets (Alcaraz and

[4] Most, but not all: Linguistic context is obviously lost and the line boundaries. This problem is being considered towards upcoming developments of handwritten search and retrieval technologies.





Number of:	Total
Pages	881
Lines	19 764
Running words	196 858
Size of lexicon	20 931
Running characters	756 122
Size of character set	77

Figure 3: "PLANTAS", XVII century Botanical Specimen Manuscript Collection of seven volumes written by a single writer in Old Spanish; page image examples and data set used for experiments on Vol. I.

WienSanktUlrich) correspond to more challenging early modern image collections exhibiting many of the difficulties entailed by medieval writing styles. The results of these laboratory experiments are presented in this section.

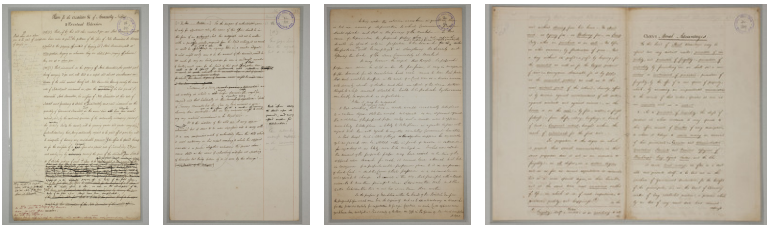
Further information regarding these data sets and the corresponding full collections can be found on the tranScriptorium web site (see section 1).

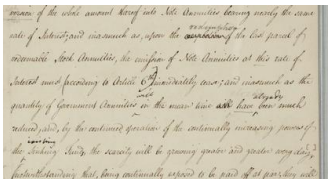
Evaluation measures

The standard *recall* and *interpolated precision* measures (Manning et al. 2008) are used to assess the effectiveness in all search experiments.

For a given query and confidence threshold, *recall* is the ratio of relevant image regions (lines) correctly retrived by the system (often called "hits") with regard to the total number of relevant regions existing in the set of test images. *Precision*, on the other hand, is the ratio of hits with regard to the number of regions retrieved (both correctly or incorrectly).

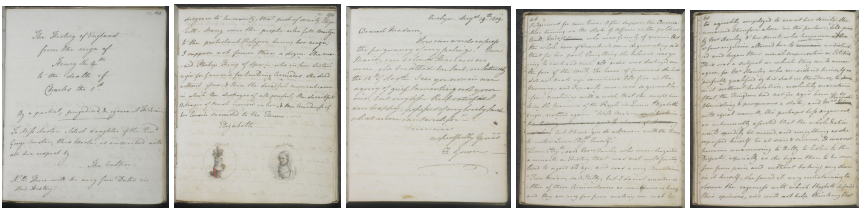
By altering the confidence threshold, different related values of recall and precision can be obtained. These values can be plotted as a *Recall-Precision* curve. In a perfect system, this curve would rise vertically from point $(1, 0)$ to $(1, 1)$ before plateauing to $(0, 1)$. Such a system should exhibit a full precison (1) independently of the confidence threshold. This would, in fact, be the behaviour of a conventional plaintext retrieval system tested on perfect transcripts of the images of the test set. A reasonable KWS

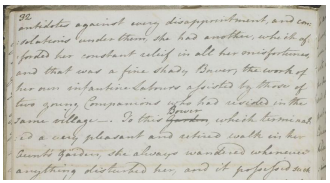




Number of:	Total
Pages	433
Lines	11 473
Running words	106 905
Size of lexicon	9 717
Running characters	550 674
Size of character set	86

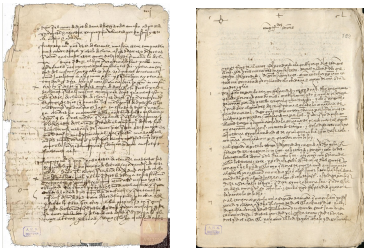
Figure 4: "BENTHAM", XVIII century collection of over 4,000 volumes of drafts and notes, written in English by several writers; page image examples and data set of 433 selected page images used in the experiments.





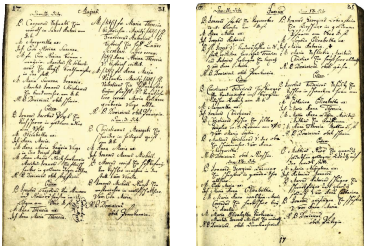
Number of:	Total
Pages	128
Lines	2 693
Running words	25 291
Size of lexicon	3 567
Running characters	118 881
Size of character set	81

Figure 5: "AUSTEN", Jane Austen's *Juvenilia*: XVIII century single hand manuscript in English; page image examples and "Volume The Third" data set used in the experiments.



- Close to 1000 page images; *miscellaneous hands, complex writing*
- About *30%* loosely *abbreviated words.*
- Experiments on 44 pages, cross-validation test
- *Lexicon & Query set*: approx. 3 400 keywords
- Training with *diplomatic transcripts*

Figure 6: "ALCARAZ", XVI century Spanish Inquisition trial against Pedro Ruiz de Alcaraz; example page images and details of the data set used in the experiments.



- Tens of thousands of two-column pages; *single hand*, but *mixed script compelex writing*
- Experiments on 52 pages, cross-validation test
- *Lexicon & Query set*: approx. 2 300 keywords
- Training with *diplomatic transcripts*

Figure 7: "WienSanktUlrich", XVI century German/Latin handwritten birth records (Wien). Example page images and details of the data set used in the experiments.

system should provide curves that rise beyond the graph's diagonal – the closer it gets to the upper right corner (point (1, 1), the better.

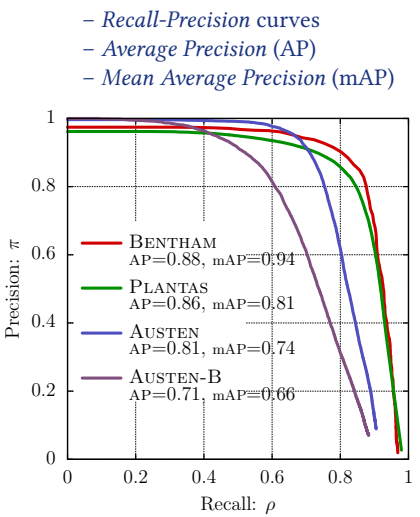
Results are also reported in terms of overall *average precision* (AP) and *mean AP* (mAP) obtained by calculating the area under the Recall-Precision curve. Both AP and mAP are popular scalar assessment measures for KWS.[5]

Results

Indexing and search results for the data sets described above are presented in figures 8 and 9. The results visualized in figure 8 correspond to the relatively modern (and less problematic) data sets (Plantas, Bentham and Austen). For the purpose of comparison, the results of figure 8 are also summarized (as gray curves) in 9, which mainly shows the results of the more challenging early modern data sets (Alcaraz and WienSanktUlrich)

In the case of the Austen data set, two experiments were carried out. In the first one, we adopted a conventional training-testing setting; i.e. KWS models were trained with annotated data of the same collection and performance was measured on an

[5] For details on these assessment measures see Toselli et al. 2016.



Data sets training and test details

- **BENTHAM:** *miscellaneous hands. Training*: 400 pages from Bentham, 87 char. HMMs, 2-gram LM trained on Bentham texts; Lexicon 9 341 tokens.
  Test: 33 pages; query set: 6 962 keywords
- **PLANTAS (VOL-I):** *single hand. Training*: 224 pages from *Plantas*, 77 char. HMMs, 2-gram LM trained with the training set + book glossary transcripts. Lexicon 11 561 tokens.
  Test: 647 pages; query set: 9 945 keywords
- **AUSTEN:** *single hand. Training*: 50 Austen pages, 81 char. HMMs, 2-gram LM trained on Austen texts; Lexicon 20K tokens.
  Test: 78 pages; query set: 2 281 keywords
- **AUSTEN-B:** *single hand.* ***No training***; using Bentham character HMMs, lexicon and LM.
  Test & query set: Same as for **AUSTEN**

Figure 8: Results on XVII-XIX century manuscript image collections

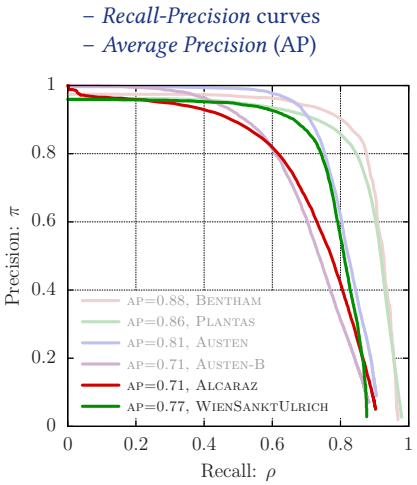
independent test set of the same collection. In the second experiment (AUSTEN-B), we used models which had been trained with transcribed BENTHAM images (the same used for the experiment with the BENTHAM data set) to index the images of the test set of the AUSTEN collection. This experiment was aimed at exploring whether a handwritten image collection can be indexed fully automatically without previous training an that particlular collection by using KWS models previously trained with images of similar handwritting styles.[6]

Good test results were achieved for all data sets. As expected, the results for the more difficult early modern data sets were less satisfying. However, even with this outcome the system can be used in practice to reliably find relevant information. The results for AUSTEN without training (i.e. using models trained for other, similar collections) were also somewhat inferior to those obtained with proper training with Austen data, but they still suffice to guarantee a successful use in practice.

Overall, the results presented above are competitive in comparison to results mentioned in the literature for classical KWS systems.[7] However, one may argue that these good laboratory results may not translate into a similarly satisfying prac-

[6] The writing style of AUSTEN was similar to the style of some of the writers of the BENTHAM collection (written in the same language and historical period).

[7] See Rath and Manmatha 2007; Rodríguez-Serrano and Perronnin 2009; Fischer et al. 2012; Frinken et al. 2012; Wshah et al. 2012; Toselli and Vidal 2013b; Toselli et al. 2016.



Data sets training and test details

- **BENTHAM:** English, *miscellaneous hands.* *Training*: 400 pages; *Query set*: 6 962 keywrds
- **PLANTAS-I:** Spanish, *single hand.* *Training*: 224 p.; *Query set*: 9 945 keywords
- **AUSTEN:** English, *single hand.* *Training*: 50 pages: *Query set*: 2 281 keywords
- **AUSTEN-B:** English, *single hand.* ***No training*** (Bentham models). *Query set*: 2 281 keywords
- **ALCARAZ:** Spanish, *multi-hand.* *Training*: 44 pages, 70 char. HMMs, 2-gram LM trained on training transcripts; Lexicon 3 405 tokens. *Test*: Cross-val.; *Query set*: 3 400 keywords
- **WIENSANKTULRICH:**German/Latin, *one hand.* *Training*: 52 pages, 74 char. HMMs, 2-gram LM from training transcripts; Lexicon 2 303 tokens. *Test*: Cross-val.; *Query set*: 2 256 keywords

Figure 9: Results on early modern collections of manuscript images.

tical search experience. Considering, for instance, the search for information in the WIENSANKTULRICH collection, the user will try to find names of persons, cities, or possibly professions. In this scenario, an operational point such as `Recall`≈ 0.7 and `Precision`≈ 0.9 (see fig. 9) would fail to retrieve an average of 30% of the lines containing the query word, while about 10% of the retrieved lines would be false hits.

Several factors, however, contribute to a search experience much more positive than would be expected from these numbers. Firstly, searching for information in manuscript images can by no means be compared to conventional information retrieval where there is no uncertainty about the query words contained in the (electronic text) documents. In manuscript images the only approach generally available nowadays is a manual search; this entails visually scanning each of the (maybe thousands or millions) page images whilst trying not to miss image regions containing the query word. Here, even an Average Precision (AP) as low as 0.5 may prove extraordinarily useful in comparison to the basis of a manual search. Secondly, the results of figure 9 are averaged for a query set of 2 256 words. This set contains every token seen in training and in the test sets, including function words and many other words (shorter, more difficult to spot) which are not usually query targets. For proper names, results

turn out generally better (but more experiments need to be conducted to objectively validate this assertion). Finally, given the precision-recall trade-off search model, the user is not expected to be content with a fixed operational point. Depending on the interest in finding only some, or most of, the occurrences of a given query word, the user will try increasing or decreasing threshold values until he or she is satisfied with the results and/or acknowledges to have met the limitations of the system.

The demonstration systems described in the next section can be used to gain first-hand experience of the capabilities of the systems in question and the significance of the results presented in this section.

4 Demonstration systems

The indexing and search engines used to obtain the results presented in section 3 are also used to support demontstration systems which can be publicly accessed online. Most of these demonstrators are available via the demonstrations section of the tranScriptoriun web site or via the following direct link:

http://transcriptorium.eu/demots/kws

It has to be be pointed out, however, that the demonstration for the PLANTAS collection does *not* include the first volume of the collection which was used in the laboratory experiments presented in the previous section.[8] In this case, the demonstrator is a working system proper, useful to find information in the about 1 000 pages of the *untranscribed* Vol. VII of the same PLANTAS collection. The optical and language models trained with pages of Vol. I and used to conduct the laboratory experiments were used to index the new, untranscribed Vol. VII fully automatically. Hence, this demonstrator can be seen as a typical and fitting example of the manifold possibilities provided by the technology presented in this paper.

5 Conclusion and outlook

A formal probabilisitic framework has been introduced for hierarchical indexing and searching large collections of handwritten documents. Empirical results with a variety of historical collections exhibiting differnet challenges and levels of complexity assess the usefulness of these methods in practice. Models trained for a given collection can provide a useful performance on images from other similar collections without need for (re-)training. Several demonstrators have been implemented and made publicly available online to allow first-hand experience in real queries.

[8] An older demonstration for Vol. I of PLANTAS is available at <http://cat.prhlt.upv.es/kws-demos>.

Future endeavors are planned to adress the following issues:

- So far, line-regions are considered the most fundamental elements to be indexed. This entails a requirement for automatic line detection and extraction. While there are fairly accurate automatic line detection techniques for textual data, results lack stability; these techniques are not stable enough to reliably tackle the significant variability in image quality and layout usually exhibited by historical manuscripts. Hence, at times a number of page images may appear in which line detection has failed drastically. As a result, these pages remain unindexed. Our current work aims at considering full page images as the lowest indexing level in an attempt to completely circumvent the line detection bottleneck.
- Techniques presented here require a predefined, possibly very large register of words to be indexed. Three approaches are currently being developped in order to overcome this limitation:
  - Probability *smoothing* techniques based on word similarities derived from character confusion probabilities
  - A *back-off* approach carrying out a computationally more extensive character-level search for queries involving non-indexed words
  - Do not longer insist in indexing *given* keywords; instead, find all the text elements which are likely to be "words" and just index all these "pseudo-words" blindly.
- All techniques and experiments described in this paper assume that a user query has the length of a single word. The development of techniques for multiple word and combined queries is currently in progress. Boolean and word sequence combinations in particular are already supported and formal evaluation results will be available in due course.

6 Acknowledgments

Thanks are due to Héctor Toselli, Joan Puigcerver and the HTR team at the PRHLT research center for their ideas and their work on the experiments illustrated in this paper. The author's work was partially supported by the Generalitat Valenciana under the Prometeo/2009/014 project grant ALMAMATER, and through the EU projects: HIMANIS (JPICH programme, Spanish grant Ref. PCIN-2015-068) and READ (Horizon-2020 programme, grant Ref. 674943).

Bibliography

Alabau, Vincent, Carlos D. Martínez-Hinarejos, Verónica Romero, and Antonio L. Lagarda. "An iterative multimodal framework for the transcription of handwritten historical documents." *Pattern Recognition Letters* 35 (2014). 195–203.

Ball, Gregory R., Sagur N. Srihari, Harish Srinivasan et al. "Segmentation-based and segmentation-free methods for spotting handwritten arabic words." *Tenth International Workshop on Frontiers in Handwriting Recognition.* 2006.

Bosch, Vicente, Alejandro Héctor Toselli, and Enrique Vidal. "Statistical Text Line Analysis in Handwritten Documents." *Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR'12)* 2012. 201–206.

Cao, Huaigu, Anurag Bhardwaj, and Venu Govindaraju. "A probabilistic method for keyword retrieval in handwritten document images." *Pattern Recognition* 42.12 (2009). 3374-3382. DOI: 10.1016/j.patcog.2009.02.003.

Fischer, Andreas, Andreas Keller, Volkmar Frinken, and Horst Bunke. "Lexicon-free handwritten word spotting using character HMMs." *Pattern Recognition Letters* 33.7 (2012). 934-942. DOI: 10.1016/j.patrec.2011.09.009.

Frinken, Volkmar, Andreas Fischer, R. Manmatha, and Horst Bunke. "A Novel Word Spotting Method Based on Recurrent Neural Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.2 (2012). 211-224. DOI: 10.1109/TPAMI.2011.113.

Graves, Alex, Marcus Liwicki, S. Fernández, Roman Bertolami et al. "A Novel Connectionist System for Unconstrained Handwriting Recognition." *IEEE Transaction on Pattern Analysis and Machine Intelligence* 31.5 (2009). 855–868.

Kamel, Ibrahim. "On Indexing Handwritten Text." *International Journal of Multimedia and Ubiquitous Engineering* 5.2 (2010).

Kolcz, Aleksander, Joshua Alspector, and Marijke F. Augusteijn. "A Line-Oriented Approach to Word Spotting in Handwritten Documents." *IEEE Transactions on Pattern Analysis & Applications* 3 (2000). 153–168. DOI: 10.1007/s100440070020.

Manmatha, Raghavan and Jamie L. Rothfeder. "A scale space approach for automatically segmenting words from historical handwritten documents." *Pattern Analysis and Machine Intelligence* 27.8 (2005). 1212–1225.

Manmatha, Raghavan, Chengfeng Han, and Edward M. Riseman. "Word Spotting: a New Approach to Indexing Handwriting." *1996 Conference on Computer Vision and Pattern Recognition (CVPR'96)*, June 18-20, 1996. San Francisco (CA): IEEE, 1996. 631–637.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (eds.). *Introduction to Information Retrieval.* New York (NY): Cambridge University Press, 2008.

Papavassiliou, Vassilis, Themos Stafylakis, Vassilis Katsouros, and George Carayannis. "Handwritten document image segmentation into text lines and words." *Pattern Recognition* 43.1 (2010). 369–377.

Puigcerver, Joan, Alejandro H. Toselli, and Enrique Vidal. "Querying out-of-vocabulary words in lexicon-based keyword spotting." *Neural Computing and Applications* (2016). 1-10. DOI: 10.1007/s00521-016-2197-8.

Rath, Tony and Raghavan Manmatha. "Word spotting for historical documents." *International Journal on Document Analysis and Recognition* 9 (2007). 139-152.

READ: *Recognition and Enrichment of Archival Documents.* <http://read.transkribus.eu>.

Rodríguez-Serrano, José A. and Florent Perronnin. "Handwritten word-spotting using hidden Markov models and universal vocabularies." *Pattern Recognition.* 42 (2009). 2106-2116. DOI: 10.1016/j.patcog.2009.02.005.

Romero, Verónica, Alejandro Héctor Toselli, and Enrique Vidal (eds.). *Multimodal Interactive Handwritten Text Recognition.* New Jersey, London, Singapore et al.: World Scientific Publishing, 2012.

Terasawa, Kengo, and Yuzuru Tanaka. "Slit Style HOG Feature for Document Image Word Spotting." *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'09) 2009.* 116–120.

Toselli, Alejandro Héctor et al. "Multimodal Interactive Transcription of Text Images." *Pattern Recognition* 43.5 (2010). 1814–1825.

Toselli, Alejandro Héctor, Enrique Vidal, and Francisco Casacuberta (eds.). *Multimodal Interactive Pattern Recognition and Applications.* London: Springer, 2011.

Toselli, Alejandro Héctor, Verónica Romero, Moisés Pastor-i-Gadea, and Enrique Vidal. "Transcribing a 17th century botanical manuscript: Longitudinal interactive transcription evaluation and ground truth production." *Digital Scholarship in the Humanities* 2017. DOI: 10.1093/llc/fqw064.

Toselli, Alejandro Héctor, and Enrique Vidal. "Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents." *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR'13)* 2013. 501–505.

Toselli, Alejandro Héctor, Enrique Vidal, Verónica Romero, and Volkmar Frinken."HMM Word Graph Based Keyword Spotting in Handwritten Document Images." *Information Sciences* 370-371 (2016). 497-518. DOI: 10.1016/j.ins.2016.07.063.

tranScriptorium. 2013-2015. <http://transcriptorium.eu>.

Vidal, Enrique, Luis Rodríguez, Francisco Casacuberta, and Ismael Garcìa-Varea. "Interactive pattern recognition." *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction.* London: Springer, 2007. 60–71.

Vidal, Enrique, Alejandro Héctor Toselli, and Joan Puigcerver. "High performance query-by-example keyword spotting using query-by-string techniques." *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR'15)* 2015. 741–745.

Vinciarelli, Alessandro, Samy Bengio, and Horst Bunke. "Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.6 (2004). 709–720.

Wshah, Safwan, Gaurav Kumar, and Venu Govindaraju. "Script Independent Word Spotting in Offline Handwritten Documents Based on Hidden Markov Models." *Proceedings of the International Conference on Frontiers in Handwriting Recognition. (ICFHR'12)* 2012. 14–19.

Tracing: A Graphical-Digital Method for Restoring Damaged Manuscripts

Dariya Rafiyenko

Abstract

Different kinds of graphical properties of manuscripts such as layout, marginalia, handwriting or text decorations are crucial for the palaeographic and philological analysis thereof. These properties help to locate the manuscript in time and space, as well as enhance the philological analysis of the text. However, in the case of ancient historical documents, this can be considerably impeded by various kinds of damages such as deterioration, erasure, moulds, fading, staining or overwriting, just to name a few. The aim of this paper is to provide a new and handy method for digital reconstruction referred to as *Tracing* that allows quite accurate reconstructing of the original graphical appearance of a damaged manuscript without requiring considerable technical expertise. *Tracing* is a non-invasive method that crucially relies on high-resolution digital images of the manuscript. Its application is illustrated here on the basis of the palimpsested manuscript *Vaticanus graecus* 73. *Tracing* was employed in order to restore the earlier, underlying text layer (*scriptio inferior*) on 12 folios or 24 pages. The results are quality images of the reconstructed manuscript pages that faithfully render the graphical properties of the original. These images may immediately be used for palaeographical and philological analyses.

Zusammenfassung

Für die paläographische und philologische Analyse von Handschriften sind mit dem Layout, den Marginalien, der Form der Handschrift oder der Textausschmückung ganz verschiedene Arten graphischer Merkmale von großer Bedeutung. Das Verständnis dieser Eigenschaften unterstützt nicht nur die Verortung von Handschriften in Zeit und Raum, sondern kommt auch der philologischen Analyse zugute. Bei manchen, besonders bei älteren Handschriften kann dies durch verschiedene Arten von Beschädigungen behindert werden: Verfall, Verblassung, Verfärbung, Ausradierung, Flecken oder Überschreibung – um nur einige zu nennen. Dieser Beitrag stellt mit der Nachzeichnung eine praktische Methode für die digitale Rekonstruktion vor, die eine getreue Nachbildung des ursprünglichen graphischen Erscheinungsbildes erlaubt ohne besondere technische Kenntnisse zu erfordern. Nachzeichnung ist ein nicht-invasives Verfahren, das entscheidend von hoch aufgelösten digitalen

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 121–135.

Abbildungen der Handschriften abhängt. Die Anwendung wird hier am Beispiel der Palimpsest-Handschrift *Vaticanus graecus* 73 vorgeführt. Die Nachzeichnung wurde hier angewandt, um den früheren, zuunterst liegenden Text (*scriptio inferior*) auf 12 folios oder 24 Seiten wiederherzustellen. Das Ergebnis sind gute Bilder der rekonstruierten Handschriftenseiten, die die graphischen Eigenschaften des Originals getreu wiedergeben. Diese Bilder können unmittelbar für die paläographische und philologische Analyse genutzt werden.

1 Introduction

Examination of historical documents is often impeded by various damages of the manuscript, for instance, by fading, staining, bleed-through, moulding, palimpsesting, and other forms of mutilation. In this paper, I primarily focus on palimpsested manuscripts that are frequently found in the Greek, Armenian, Georgian or Syriac traditions (Maniaci 2015, 73) and present a particularly complex case of damage: in the process of palimpsesting, the text is intentionally scraped or washed off so as to reuse the pages for copying new texts (Thomson 1912, 64–66). Needless to say, damages resulting from palimpsesting considerably decrease the legibility of the original text and make the graphical appearance of the manuscript pages no longer readily available for any kind of study.

Palimpsests have been paid much attention in modern paleographic research since many of them contain texts from earlier time periods that are otherwise unknown (a list of such manuscripts can be found in, *inter alia,* Wattenbach 1896, 299–317 and Thompson 1912, 65–66). An increasingly strong interest in palimpsests started to arise in the eighteenth century. This is also the time when a number of methods to uncover the underlying layers of writing in palimpsests were invented and adopted (Dillon 2007, 16–22). First of all, invasive methods were applied. These methods aimed at improving the legibility of the faded text by the use of chemical reagents such as oak-gall tincture, liver or sulphur tinctures, or Giobert tincture, a weak acid solution of potassium hexacyanoferrate (II) (Albrecht 2015, 31). As this often caused serious damaging effects, invasive methods were largely abandoned at the beginning of the twentieth century whereas various sparing optical techniques started to be primarily used for decipherment purposes. The documents were examined or imaged by means of light at different – also invisible to the human eye-wavelengths. Ultraviolet light was commonly used during the twentieth century. Nowadays, a number of more advanced digital imaging techniques are successfully applied to decipher the underwritings of palimpsests. As it is not possible to give a thorough overview of all such techniques here, I refer to the recent short description thereof in Albrecht (2015,

26–27, 31–33) with further references therein; below I mention some recent projects implementing such kind of techniques.

When it comes to the study of palimpsests, the main concern has always been to retrieve as much information of the *scriptio inferior* as possible, whereas little or no attempt has been made to restore the original look of the underlying document and make it available for research. This might be due to a number of reasons. First, it was technically challenging to produce and publish such images in the predigital era. Secondly, the restoration of the original appearance of the document was not on the agenda of traditional philology, whose main focus was always to decipher the text and make it legible for interpretative research. As a consequence, a transcript of the deciphered text was considered as sufficient.

One may wonder why it would be so important to re-create the original look of the underlying document at all. The idea that medieval texts have to be studied along with their material representation as a single phenomenon came up with the rise of the so-called 'new' or 'material' philology in the last quarter of the twentieth century (for an overview of this editorial school, see Baker 2010, especially 440–444, and Driscoll 2010, 90–95) and emphasized once again recently in, for example, Agapitos (2008), Pierazzo and Stokes (2011), and Pierazzo (2014). One of the premises of the material philology is formulated in Driscoll as follows: "[l]iterary works do not exist independently of their material embodiments and the physical form of the text is an integral part of its meaning" (2010, 90). The text of a historical document along with its graphical appearance returns into its original context and material reality that surrounded it; and thus it can be analyzed as an intellectual and cultural artefact of its time. Consequently, the restoration of the graphical appearance of historical documents was called for. Below I will show how the analysis of the graphical appearance of the manuscript can enhance our understanding of the ways the text could have been used by the reader on the basis of the example of codex *Vat. gr.* 73. The method to be envisaged below allowed me to determine a system of marginalia and pictograms in the margins of this manuscript. These are intended to help the reader to navigate through the content – a phenomenon that was ignored by the previous editions altogether.

Furthermore, graphical properties of the manuscript are crucial for the paleographic research, decipherment or dating. Properties of the original text composition such as punctuation may help to uncover the syntactic structure of the sentences as well as the functions of particular expressions within the sentences. Last but not least, the graphical appearance of the text may have impact on the philological interpretation of the text. For example, the layout of the text can provide clues as to its segmentation into chapters, sections, passages or similar. I conclude that the analysis of the graphical appearance of the manuscript is an indispensable part of the investigation of the manuscript.

When it comes to the facsimile edition of a historical document, there are three main options as to how it may be produced: (1) *reproduction*, (2) *restoration*, and (3) *reconstruction*. Reproduction (1) is a type of facsimile edition in which the photographical images are published as they are with no image processing. This approach is suitable in cases where the text – perhaps despite some minor damages – is immediately visible and legible. One of the earliest examples of this approach is, for instance, the edition of *Wulfila's Gothic Bible* by Hans Henning, published as early as 1913.

Both restoration (2) and reconstruction (3) presuppose that the images of the manuscript are published after some image processing to increase the legibility of the damaged text. During the process of restoration (2) particular characteristics of the original images of the manuscript are adjusted so as to enhance the legibility of the respective textual layer. Restoration (2) is a widely used technique. It was applied, for instance, within the *Digital Image Archive of Medieval Music* (DIAMM) project. The image processing techniques of this project, such as the global and single-area level adjustment, are described by Craig-McFeely and Lock (2006) in their *Digital Restoration Workbook*. Another restoration approach was adopted by Sparavigna (2009) while restoring Da Vinci's sketches. This approach heavily relies on manipulating colour-channels. Yet another technique similar to the former one was put forward by Stokes (2011) and adapted by Voth (2014). It is primarily designed for palimpsested manuscripts and was used to examine the oldest extant Old English manuscript on medical remedies. Technically sophisticated restoration methods relying on the combination of imaging under special conditions and image processing were applied in the projects supported by the *Early Manuscripts Electronic Library* (EMEL). The list of relevant projects can be found on the website of the EMEL project. The restored palimpsested text of Archimedes published in Easton et al. (2003), Easton and Knox (2004), Netz and Noel (2007) provides an illustration of the application of this method. Still another approach to restoration is the digital 3-D modeling of physical objects; the advantages of this approach are discussed in Brown and Seales (2001).

In the process of reconstruction (3) – as opposed to the restoration – a new graphical object is created, which should be as close to the original as possible. For example, reconstruction of the fragment of page 46 from *Vat. gr.* 73 is found in the edition of Mai (1827, 1). Its goal was to illustrate the layout and its function in the structure of the manuscript (*fig. 1* below). Another example is Gurtmann (2012). Here, an extensive digital reconstruction allowed to reveal the complete underlying text of parchment manuscript of Qur'an from Sanaa from ca. 650 CE.[1] Finally, another technique of reconstruction similar to the method of *Tracing* illustrated in Section 2 below is found

[1] This graphical reconstruction is to be published in the Brill series *Documenta Coranica* (ed. by F. Déroche, M. Marx, A. Neuwirth and Ch. Robin).

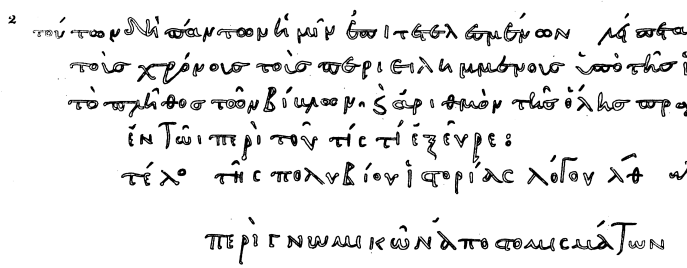


Figure 1: Reconstruction of the fragment of page 46 from *Vat. gr.* 73 published in the edition of Mai (1827, 1).

in Butcher and Hrynick (2012). Here, within the scope of the O*xford Outremer Map* project, a map from thirteenth century was digitally reconstructed, making once barely legible writings and images clearly visible.

Under both approaches, the restoration (2) and reconstruction (3), the critical question is as to how many amendments – if at all – may be made on images. Ideally, it is the duty of the researcher to ensure that facts are not manipulated and only most plausible emendations are made. Discussion of the ethical side of image manipulation may be found in Craig-McFeely and Lock (2006, 35–36, 53–54), Craig-McFeely (2008, §62), and Stokes (2011, 20). One possible solution to this is to supply the publication with the original images and the full list of all manipulations made.

2 The manuscript

In the next section, the technique of *Tracing* is outlined on the basis of *Vat. gr.* 73, a palimpsested parchment manuscript preserved in *Biblioteca Apostolica Vaticana*, Vatican City. Its upper textual layer, or *scirptio superior*, is dated back to the fourteenth century and contains the speeches of Aelius Aristides and the dialog *Gorgias* of Plato (Mercati and de'Cavalieri 1923, 67). It was only in the first quarter of the nineteenth century that an earlier textual layer, or *scriptio inferior*, was discovered by Angelo Mai (1782–1854), a celebrated philologist of the nineteenth century famous for finding a great number of hitherto unknown palimpsested texts of ancient authors (Dillon 2007, 10–22). He identified the *scriptio inferior* as one of the volumes of the Constantinian excerpt collection, or *Excerpta historica Constantiniana*, a tenth century historic encyclopaedia written in Constantinople in Ancient Greek language (Németh 2010). On the basis of palaeographical and codicological characteristics Jean Irigoin (1959) dated the manuscript into the first half of the tenth century, assuming that the *Vat. gr.* 73 is the original volume of the *Excerpta Constantiniana* assembled on behalf of

the emperor Constantine VII (913–959) for the imperial library. Only a subset of the original leaves of the *Excerpta Constantiniana* were palimpsested in the fourteenth century. It is assumed that the 177 folios, or 354 pages, preserved until today constitute around two third of the original manuscript.

Physical dimensions of the manuscript are 350/355 × 270/275 mm; writing surface covers approximately 255/260 × 185/200 mm. Written in 32 lines per page with approximaly 45–55 characters per line, the letters of both layers are about 5 mm high. The *scriptio superior* is written immediately above the *scriptio inferior*, fully covering it and extremely reducing its legibility (see fig. 2).

Rafiyenko (forthcoming) represents the reconstruction of the following 24 pages of *scriptio inferior*: 301, 302, 349, 350, 203, 204, 205, 206, 343, 344, 299, 300, 261, 262, 337, 338, 309, 310, 323, 324, 327, 328, 275 and 276.[2] These pages contain excerpts from an anonymous historiographer, the so-called *Anonymous post Dionem*, oftentimes identified as Peter the Patricius, an official and ambassador from the time of Justinian I (527–565)(Antonopoulos 1990). The text of the *Anonymous* is an account of Rome's history from the reign of Augustus (27 BCE–14 CE) to Constantine the Great (306–337 CE).

The text of the *scriptio inferior* of the *Vat. gr.* 73 has been edited twice: parts of it were edited by Mai in 1827; the full text was edited by Boissevain in 1906. Both editors studied the manuscript in autopsy and both of them used chemicals in order to enhance the legibility of the lower text (Mai 1827, XXXI–XXXIII; Boissevain 1884, 25). However, chemical treatments can considerably deteriorate the preservation condition of a palimpsest with the lapse of time (Wattenbach 1896, 311–312). In the case of the *Vat. gr.* 73, it remains unclear to what extent the manuscript was treated by Mai and Boissevain and how the treatment affected its condition. According to my own assessment, the legibility of the lower text did not considerably change since then. Previous editors were able to decipher most parts of the text (the editions of Mai 1827 and Boissevain 1906 contain almost no gaps in the text). Currently, the amount of the lower text which can be discerned with the naked eye amounts up to 90–95%.[3] The ink is for the most part discernible with the naked eye. However, the degree of preservation varies significantly from page to page, from line to line, and even from mark to mark. A number of images from *Vat. gr.* 73 are contained in Németh (2015).

---

[2] The pagination in the *Vat. gr.* 73 has two peculiarities. First, page numbers instead of folio numbers are traditionally used for reference in *Vat. gr.* 73. Secondly, the pagination reflects the sequence of the pages in the palimpsested manuscript and therefore becomes re-ordered when the page sequence of the original manuscript is reconstructed.

[3] According to my own experience from the study of the manuscript both in the autopsy and by means of high-resolution digital images.

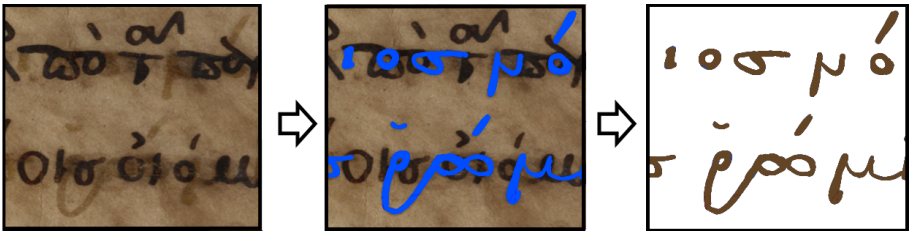


Figure 2: The process of the graphical reconstruction of the *Vat. gr.* 73 (fragment of p. 301)

3 The method

The impetus to develop the method of reconstruction of the *Vat. gr.* 73 came from the wish to facilitate the process of the autopsy of the manuscript for a new edition (Rafiyenko, forthcoming). Deciphering such a damaged text revealed itself as labour-intensive and, hence, time-consuming work. Collating took sometimes up to twenty hours per page. Nonetheless, irrespective of the time invested, previous collation brought little when a passage from the manuscript had to be consulted repeatedly; and subsequent revisions were almost as time-consuming. This called for a different method of decipherment that would allow fixing the deciphered characters in a digital form. The resulting images revealed themselves as a clear copy of the lower text and its original graphic appearance (cf. fig. 2 and fig. 3 below). The essence of the technique lies in manual re-tracing and re-drawing the contours with the stylus on the touch screen on significantly enlarged images. This allows rendering the scriber's handwriting very close to the original (cf. fig. 2).

3.1 Technical requirements

An image processing software with the Brush Function such as Paint.NET, GIMP, Photoshop, ImageJ or other is sufficient. Furthermore, one needs a digital drawing pad or drawing tablet and, finally, high resolution photos of a manuscript for the reconstruction. It is advisable to have a monitor with high quality resolution.

3.2 Tracing

The images are drawn with digital painting technique in an image-processing application. In order to be able to separate the original images from the reconstruction, the latter are drawn on a separate layer positioned above the layer containing the original image of the manuscript. Magnification ensures high accuracy of imitation of the ductus and of the characteristic shapes of the ink marks.

In my case, a magnification of eight to ten times has proved itself as optimal. For this purpose, I used a Hewlett Packard notebook with resistive touchscreen. The images were created with the Brush Function in the Paint.NET image-processing application. In a non-digital environment, one could potentially achieve comparable results by putting a transparent slide upon the image and, subsequently, re-drawing the ink marks on the slide manually. However, the crucial advantage of the digital method here is the possibility of modifying the characteristics of the original image, which allows to discern and to trace the original ink marks with a higher level of fidelity.

3.3 Results

The resulting images can be characterized as a two-dimensional, exact and truthful representation of the manuscript's underwritings. They represent the surface of the lower text in terms of a *topographical* edition[4] (the term coined by P. Sahle in personal communication).

In Rafiyenko (forthcoming), the exact appearance of the original manuscript pages – not readily discernible behind the ink marks from the fourteenth century layer – is restored (see fig. 3). The high level of granularity allows determining the ductus and the characteristic shapes of ink marks in all parts of the lower text. Thus, maximum fidelity to the features of the handwriting is achieved and such properties as the colour of the ink or spatial positioning are straightforwardly reproduced.

As the image of page 302 from the *Vat. gr.* 73 shows (fig. 3), graphical reconstruction of the manuscript page gives a clear picture of its overall appearance before it was palimpsested. It offers a number of advantages. First, it allows a better understanding of the layout of the *Vat. gr.* 73. It is clearly visible that the initial letter ὅ (see lines 1, 5, 9, 12, 16, 27, 29 and 30 of page 302 on fig. 3) is used as a visual marker of the starting point of each new excerpt in the *Excerpta Constantiniana,* being set off by the space left blank before it and by the use of the reddish ink. The visual appearance of initials is important here as it highlights the logical structure of the text and shows that borders of each excerpt were clearly marked as well as that the excerpt itself was construed as the smallest single unit of the text structure.

Furthermore, marginalia and pictograms in *Vat. gr.* 73 can now be studied since their exact positioning and design are clearly visible in the graphical reconstruction. On page 302, there are two marginalia (placed opposite the lines 1 und 5, see fig. 3) and seven pictograms in different state of preservation (placed on the left margin opposite the lines 7, 9, 14, 18, 26, 30 and 32, see fig. 3). The palaeographical characteristics thereof unequivocally indicate that both marginalia and pictograms belong to the hand

[4] *Topographical edition* refers to any edition which is a two-dimensional representation of the surface of the original document that was created by means of reconstruction.

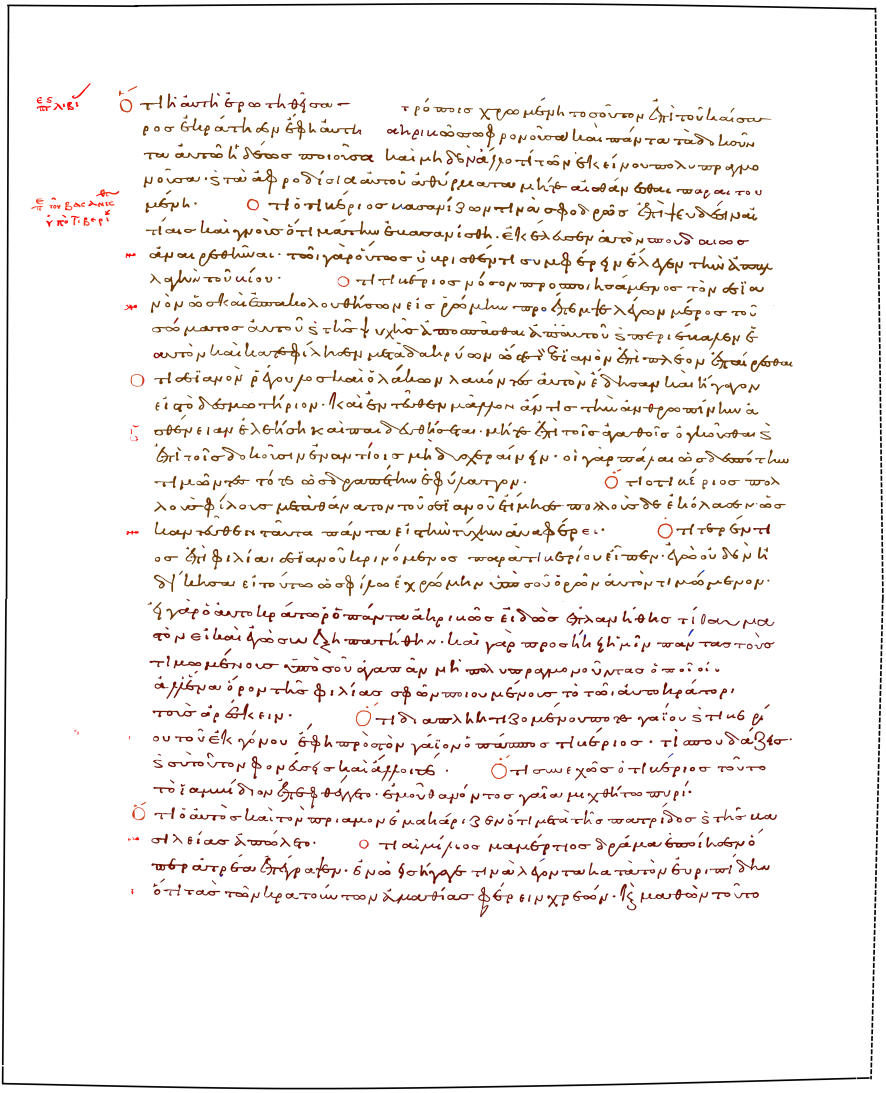


Figure 3: Graphical reconstruction of page 302 from the *Vat. gr.* 73

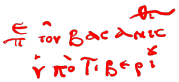
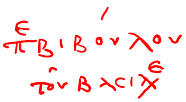
		
p. 302,1	p. 302,5–6	p. 328,9–10

Table 1: Samples of marginalia from the *Vat. gr.* 73.

of the main scribe and thus were designed in the tenth century by the compilers of the manuscript. Marginalia, 32 examples of which can be found on the 24 investigated pages of the manuscript (in Rafiyenko, forthcoming), were written in red ink and placed on the outer margins of the manuscript. They are positioned consistently at the beginning of each excerpt and indicate an acting person. As many of them are pointing to Roman emperors, it may be the case that they were also used as the chronological labels in the text. Pictograms, 147 samples of which are found on the 24 pages of the manuscript, were drawn with the same red ink as the marginalia and placed coherently on the left side of the main text. They are positioned in the middle or at the end of an excerpt and indicate the most important phrase of a given excerpt, its essence.

The reconstruction allows to establish different forms of pictograms (see *table 2*). The function of most of them could be discerned. For example, form (2) refers to passages with explicitely ironical intent, form (4) to citations of ancient authors in the text. The most numerous form (1) was presumably used without special function because it refers to a great number of passages which cannot be easily categorized (pictograms with similar function are found in papyri from Egypt of the period from 2 BCE to 7 CE, see McNamee 1992, 8).

It may be concluded that both marginalia and pictograms represent a system of content-related references that were designed to facilitate the navigation through the text of the *Excerpta Constantiniana*. The graphical reconstruction of the manuscript by means of *Tracing* makes it possible to compare the marginalia and pictograms and palaeographically analyse them (cf. *table 1*, *table 2*). The exact positioning of the marginalia and pictograms in the manuscript enhance the philological analysis of the text.

Another example of how reconstructed images can be used for the palaeographic research is presented in *table 3* below. Here, samples of the variants of the letter *epsilon* and its combinations with other letters are given. Such collations of scribal variants are important for further work with the manuscript. Notably, without the reconstruction, it is nearly impossible to acquire clear sample images of scribal variants in *Vat. gr.* 73.

As regards the truthfullness of the reconstructed images by means of *Tracing*, the major principle here may be formulated as follows: the reconstruction is based either

Nr.	Form	Number of samples in the manuscript	Indication of location in the manuscript (page, line)
(1)		86	*Passim*
(2)		13	302,26; 302,32; 206,23 etc.
(3)		2	310,25; 310,29
(4)		5	204,11; 206,2; 276,21; 276,22; 276,23

Table 2: Samples of pictograms from the *Vat. gr.* 73.

ε	-εί-	-εῖν	-σχεῖν	ἐπὶ
				
ἐτρ-	δὲ	ὥστε	ἐγὼ	ἐκ
				

Table 3: Variants of *epsilon* and its combinations with other letters in the *Vat. gr.* 73.

on the documental evidence for a character or on the unambigiously attested rests thereof. In turn, in case of ambiguity, when the form of the letter is almost entirely obscured the process of reconstruction is subject to the scholar's interpretation. In certain cases, a particular interpretation is strongly favourable because of the palaeographic norms and good acquaintance with the scriber's handwriting in this manuscript. Importantly, all amendments of this type should be marked in the critical apparatus. In other cases, where no reliable restoration can be made the space should remain blank in the reconstructed version and marked as such (cf. the marked spaces in figure 4). This also should be documented in the critical apparatus.

Certainly, these principles do not entirely exclude the possibility of over-interpretation on the part of the editor. However, the advocated method provides a much safer reconstruction tool than the traditional editorial one. Thus, in the edition of the *Excerpta Constantiniana* of Boissevain (1906), apart from tacitly made corrections of the text, the most prominent evidence of misrepresentation of the text is probably the fact that the marginalia and pictograms are neither mentioned nor represented, even though they are crucial for the understanding of the text as has been layed out above. In turn, the reconstruction of the *Vat. gr.* 73 by means of *Tracing* allows the researcher to see their exact positions, forms and the ductus and presents a more reliable source for the study of the text.

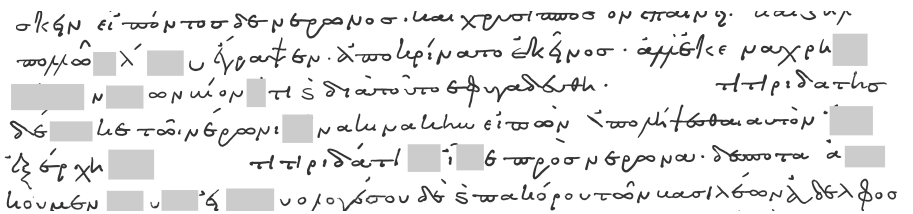


Figure 4: Illegible passages in the *Vat. gr.* 73 (p. 343, 8–13).

The method of *Tracing* has its limitations. It is time-consuming and requires a lot of effort on the part of the editor. It is, furthermore, applicable only to those manuscripts in which the lower or the damaged text may be perceived with naked eye – unless the photos have been additionally processed to enhance the legibility. Thus, *Tracing* can be considered as supplementary to more technically advanced methods such as methods relying on *multispectral imaging* (a list of projects using this technique can be found on the website of the EMEL project), *hyperspectral transformation imaging* (see Shiel, Rehbein and Keating 2009), and various techniques of image processing. Moreover, *Tracing* can also be applied in those cases in which more than one image is used as the basis for the reconstruction.

Crucially, the method has a number of advantages. Being fairly simple, it is immediately accessible to any researcher as it requires neither special software nor any technical expertise beyond the basics. At the same time, it ensures the results that cannot be achieved by simple, non-digital re-drawing. Furthermore, in contrast to automated reconstruction, the editor has the full control over the process of reconstruction here, thereby avoiding misinterpretations or mistakes made by software.

It is also advantageous concerning the philological accuracy and falsifiability of the reconstruction. This method allows the documentation of what exactly the editor sees in the lower text and what has been amended by the editor on the basis of contextual plausibility. It ensures more transparency in the process of text transcription and critical editing. Subjective decisions of the editor can be better controlled for and the requirement of falsifiability of research is more strongly obeyed than in the traditional approach.

4 Conclusion

In this paper I presented the method for reconstructing damaged manuscripts referred to as *Tracing*. It crucially relies on the re-drawing of poorly discernable contours of the original image under multiple magnification of the original size. The application

of the method was demonstrated on the basis of the palimpsested manuscript *Vat. gr.* 73. The method has a number of advantages in contrast to the traditional method. In particular, *Tracing* may be especially helpful in palaeographical and philological research because it yields qualitive pictures of the original graphical appearance of damaged manuscripts. It thus provides good empirical basis for further research on the manuscript. *Tracing* is both feasible and advantageous for the scholars of different philological subdisciplines because it does not require any advanced technical expertise nor does it require any specific technical equipment or software. Last but not least, the images produced via *Tracing* represent the editor's own artwork and, hence, should not require a copyright permission from the owning library.

Any kind of reconstruction can be considered as a step away from the real, imperfect characters of the manuscript towards their original form (as written by the scriber). While *Tracing* is about light and manual reconstruction of each and every single symbol (signs, letters, etc.) there are other, more powerful methods available that may supplement the result achieved by *Tracing.* Thus, the next step towards reconstructing transcriptions may rely on regularization of the text with the help of a font that imitates the form and the positioning of the original handwritten characters (see such an attempt in Vorbach 2012), a font where each glyph ideally would be an average of all real representations of a given character. Even though this type of reconstructing transcription diminishes the individuality of the handwritten text, it retains the topografical dimension and makes the text searchable by the computer.

Bibliography

Agapitos, Panagiotis. "Literary criticism." In Jeffreys, Elizabeth. (ed.). *The Oxford handbook of Byzantine studies.* Oxford: Oxford University Press, 2008. 77–85.

Albrecht, Felix. "Methods in palimpsest research." In Bausi, Alessandro et al. (eds.). *Comparative Oriental Manuscript Studies. An Introduction.* Hamburg: Tredition, 2015. 31–33.

Antonopoulos, Panayotis T. *Πέτρος Πατρίκιος. Ὁ Βυζαντινὸς διπλωμάτης, ἀξιωματούχος καὶ συγγραφέας* [*Peter the Patrician. The Byzantine Diplomat, Official and Author*]. (=Historical Monographs 7). Athens, 1990.

APP: *Archimedes Palimpsest Project.* <http://archimedespalimpsest.org>.

Baker, Craig. "Editing Medieval Texts." In Classen, Albrecht. (ed.). *Handbook of Medieval Studies. Terms – Methods – Trends.* Berlin: de Gruyter, 2010. 427–450.

Boissevain, Ursul Philipp. "De Excerptis Planudeis et Constantinianis ab Ang. Maio editis quae vulgo Cassio Dioni attribuntur." *Progr. Gymnasii Erasmiani* 1884–1885. Rotterdam, 1884. 13–40.

Boissevain, Urusul Philipp. *Excerpta historica iussu Constantini Porphyrogeniti confecta.* Vol. 4: *Excerpta de sententiis.* Berlin: Weidmann, 1906.

Brown, Michael S., and W. Brent Seales. "The Digital Atheneum: New Approaches for Preserving, Restoring and Analyzing Damaged Manuscripts." *Presented at ACM/IEEE-CS Joint*

Conference on Digital Libraries 2001.

Butcher, Rachel and Tobias Hrynick. "Digitally Enhancing the Map." 2012. <https://medievalomeka.ace.fordham.edu/exhibits/show/oxford-outremer-map/cleaning-the-map>.

CCP: *Corpus Coranicum Project.* <http://www.coranica.de>.

Craig-McFeely, Julia. "Digital Image Archive of Medieval Music: The evolution of a digital resource." *Digital Medievalist* 3 (2008). DOI: 10.16995/dm.16.

Craig-McFeely, Julia and Alan Lock. *Digital Restoration Workbook.* 2006. <http://www.diamm.ac.uk/redist/pdf/RestorationWorkbook.pdf>.

DIAMM: *Digital Image Archive of Medieval Music Project.* <http://www.diamm.ac.uk>

Dillon, Sarah. *The Palimpsest: Literature, Criticism, Theory.* London, New York (NY): Continuum, 2007.

Driscoll, Matthew "The Words on the Page: Thoughts on Philology, Old and New." In Quinn, Judith, and Emily Lethbridge (eds.). *Creating the Medieval Saga: Versions, Variability and Editorial Interpretations of Old Norse Saga Literature.* Odense: Syddansk Universitetsforlag, 2010. 85–104.

EMEL: *Early Manuscripts Electronic Library project.* <http://emel-library.org>

Easton Jr., Roger L., Keith T. Knox, and William A. Christens-Barry. "Multispectral Imaging of the Archimedes Palimpsest." *Proceedings of the 32nd Annual Conference on Applied Imagery Pattern (AIPR)*, 2003. 111–118.

Easton Jr., Roger L. and Keith T. Knox. "Digital Restoration of Erased and Damaged Manuscripts." In Gensler, Elana and Joan Biella (eds.). *Proceedings of the 39th Annual Convention of the Association of Jewish Libraries.* New York: Association of Jewish Libraries, 2004. <http://databases.jewishlibraries.org/node/17573>.

GIMP: GNU *Image Manipulation Program.* <http://www.gimp.org>.

Gurtmann, Hadiya. "A Qur'an written over the Qur'an – why making the effort?" *Centre for the Study of Manuscript Cultures* (CSMC), 2012. <http://www.manuscript-cultures.uni-hamburg.de/mom/2012_01_mom_e.html>.

Henning, Hans. *Der Wulfila der Bibliotheca Augusta zu Wolfenbüttel: (Codex Carolinus).* Braunschweig: C.E. Behrens, 1913.

ImageJ: *Image Processing and Analysis in Java.* <https://imagej.nih.gov/ij>.

Mai, Angelo. *Scriptorum veterum nova collectio e Vaticanis codicibus edita.* Vol. 2: *Historicorum Geacorum partes novas complectens.* Rome: Typis Vaticanis, 1827.

Maniaci, Marilena. "Parchment." In Bausi, Alessandro et al. (eds.). *Comparative Oriental Manuscript Studies. An Introduction.* Hamburg: Tredition, 2015. 72–73.

McNamee, Kathleen. *Sigla and Select Marginalia in Grek Literary Papyri.* Bruxelles: Fondation Égyptologique Reine Élisabeth, 1992.

Mercati, Giovanni, and Pius Franchi de'Cavalieri. *Codices Vaticani Graeci.* Vol. 1: *Codices* 1–329. Rome: Bibliotheca Vaticana, 1923.

Németh, András. *Imperial Systematization of the Past. Emperor Constantine VII and His Historical Excerpts.* Working copy of the doctoral thesis. Budapest: Central European University, 2010. <http://www.etd.ceu.hu/2010/mphnea01.pdf>.

Németh, András. "Layers of Restorations: *Vat. Gr.* 73 transformed in the tenth, fourteenth, and

nineteenth centuries." *Miscellanea Bibliothecae Apostolicae Vaticanae* XXI (2015). 281–330.
Netz, Revile, and William Noel. *The Archimedes Codex: Revealing the Secrets of the World's Greatest Palimpsest.* London: Weidenfeld and Nicolson, 2007.
OOM: *Oxford Outremer Map project.* <https://medievaldigital.ace.fordham.edu/mapping-projects/oxford-outremer-map-project/>.
Paint.NET: *Free image and photo editing software.* <www.getpaint.net>.
Pierazzo, Elena and Peter A. Stokes. "Putting the Text back into Context: A Codicological Approach to Manuscript Transcription." *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age* 2. Norderstedt: Books on Demand, 2011. 397–429.
Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods.* 2014. <http://hal.univ-grenoble-alpes.fr/hal-01182162/document>.
Rafiyenko, Dariya. *Konstantinische Exzerptensammlung und der Anonymus post Dionem: Begleitende Studien und Edition.* Dissertation at University of Cologne, forthcoming.
Shiel, Patrick, Malte Rehbein, and John Keating. "The Ghost in the Manuscript: Hyperspectral Text Recovery and Segmentation." *Kodikologie und Paläographie im digitalen Zeitalter-Codicology and Palaeography in the Digital Age* 1. Norderstedt: Books on Demand, 2009. 159–174.
Sparavigna, Amelia. "The Digital Restoration of Da Vinci's Sketches." 2009. <https://arxiv.org/ftp/arxiv/papers/0903/0903.1448.pdf>.
Stokes, Peter A. "Recovering Anglo-Saxon Erasures: Some Questions, Tools and Techniques." In Carruthers, Leo, Raeleen Chai-Elsholz, and Tatjana Silec (eds.). *Palimpsests and the Literary Imagination of Medieval England.* New York (NY): Palgrave Macmillan, 2011. 35–60.
Thompson, Edward Maunde. *An introduction to Greek and Latin palaeography.* Oxford: Clarendon Press, 1912.
Vorbach, Paul. *Erstellung von TrueType-Fonts zu historischen Manuskripten.* Bachelorarbeit. Julius-Maximilians-Universität Würzburg, Institut für Informatik, Lehrstuhl für Informatik II. Würzburg, 2012.
Voth, Christine. "What lies beneath? The application of digital technology to uncover writing obscured by a chemical reagent." *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age* 3. Norderstedt: Books on Demand, 2014.
Wattenbach, Wilhelm. *Das Schriftwesen im Mittelalter.* 3. Aufl. Leipzig: Hirzel, 1912.

Automatable Annotations – Image Processing and Machine Learning for Script in 3D and 2D with GigaMesh

Bartosz Bogacz, Hubert Mara

Abstract

Libraries, archives and museums hold vast numbers of objects with script in 3D such as inscriptions, coins, and seals, which provide valuable insights into the history of humanity. Cuneiform tablets in particular provide access to information on more than three millennia BC. Since these clay tablets require an extensive examination for transcription, we developed the modular GigaMesh software framework to provide high-contrast visualization of tablets captured with 3D acquisiton techniques. This framework was extended to provide digital drawings exported as XML-based Scalable Vector Graphics (SVG), which are the fundamental input of our approach inspired by machine-learning techniques based on the principle of word spotting. This results in a versatile symbol-spotting algorithm to retrieve graphical elements from drawings enabling automated annotations. Through data homogenization, we achieve compatibility to digitally born manual drawings, as well as to retro-digitized drawings. The latter are found in large Open Access databases, e.g. provided by the Cuneiform Database Library Initiative (CDLI). Ongoing and future work concerns the adaptation of filtering and graphical query techniques for two-dimensional raster images widely used within Digital Humanities research.

Zusammenfassung

Bibliotheken, Archive und Museen besitzen große Mengen an Objekten mit Schrift in 3D, wie z.B. Inschriften, Münzen und Siegelabdrücke. Diese erlauben wertvolle Einblicke in die Geschichte der Menschheit. Das gilt besonders für Keilschrifttafeln, die Informationen über dreieinhalb Jahrtausende vor der Geburt Christi übertragen. Weil diese Tontafeln eine gründliche Untersuchung, Umzeichung und Umschrift benötigen, haben wir das modulare *GigaMesh Software Framework* entwickelt, das eine kontrastreiche Darstellung von 3D-vermessenen Tafeln in hoher Auflösung ermöglicht. GigaMesh bietet dazu die Möglichkeit zum Export von Vektorzeichnungen im XML-basierten Scalable Vector Graphics (SVG) Dateiformat. Diese Dateien stellen die Datenbasis für Verfahren aus dem Bereich des Machine Learning dar, die wiederum auf dem Prinzip des Word Spotting beruhen. Daraus ist eine graphische

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 137–149.

Suchmöglichkeit von Symbolen bzw. Zeichen entstanden, mit der eine automatische Annotation möglich wird. Durch die Homogenisierung von Dateiformaten konnten wir eine Kompatibilität mit weiteren Quellen in Form von digital erstellten Handzeichnungen und Retro-Digitalisaten erreichen. Letztere stehen online per Open Access z.B. im Rahmen der Cuneiform Database Library Initiative (CDLI) zur Verfügung. Laufende und künftige Arbeiten sind die Adaption unserer graphischen Verfahren für zweidimensionale Rasterbilder, wie sie in den Digital Humanities häufig zu finden sind.

1 Introduction

The analysis of historical texts begins with the analysis of a document as an object. Therefore, any Digital Humanities (DH) project has its roots in digitized documents which are often represented by images consisting of a regular grid of colored pixels. These images are typically gathered using a flatbed scanner or digital photo cameras (Effinger et al. 2003). The latter are often combined with minimalistic 3D acquisiton using a laser-line to remove distortions such as bent pages of an open book. A well-known setup is known as the *Grazer Buchtisch*, which was invented by the engineer Manfred Mayer within a project of the University Library of Graz in Austria.

Other optical imaging methods capture even more information on the materiality of an object using 3D acquisiton. These systems are used ever more frequently in many disciplines within the Humanities due to increasing image resolution and decreasing costs of purchase. Especially in the field of Archaeology, a photogrammetric approach known as *Structure from Motion* (SfM) (Ullman 1979) is widely applied to simple objects such as ceramics (Mara and Portl 2013), coins (Boss et al. 2012) as well as more complex inscriptions (Krömker 2013) based on the principles of structured light and stereo analysis (Sablatnig and Menard 1992). However, there are many other means of 3D acquisiton such as *Reflectance Transformation Imaging* (RTI) (Woodham 1980) and the *KU Leuven Dome* (Willems et al. 2005).

All those metal, stone, or clay objects play an important role for research within the Humanities because they are comparatively robust by design and can transport information over long periods of time. These artifacts are well preserved and their content, i.e. the text on their surface, can be read by illuminating the surface using a light source to show characters as shadows on a bright background. Therefore, at first glance, photography appears to be a reasonable choice for documentation. However, a photo provides only one projection using one position of the light source. Furthermore, the surface of an object can have an arbitrary color (e.g. due to stains) camouflaging the *Script in 3D*. Even for relatively well-preserved objects, the information represented geometrically can become difficult to grasp.

Generalizing the challenges posed by surfaces weathered, worn, or otherwise damaged, we have to capture the geometry of an object in order to then remove the camouflaging colors in a first step. In a second step, the traces intentional left by a human being have to be illustrated using images without illumination which show meaningful features via color contrast. Such surface features can be determined by computing local curvature measures (Bertrand et al. 1848) like the *Gaussian curvature* (Gauss et al. 2007) which separates concave and convex areas. A second important measure is the mean curvature which can be used to determine the smoothness or roughness of an object area, e.g. to separate patterns of fracture from those intentionally left by craftspeople.

Measuring local curvature on surfaces is done in principle similarly to filtering raster images for which a multitude of edge detectors (or filter operators) was developed during the last five decades of Computer Science reserach, starting with the *Roberts-Cross-Operator* (Roberts 1963). In essence, those filters assume an image as a height map (cf. Digital Terrain Model) in which each gray-value of a pixel corresponds to height. By computing changes of heights in local environments, these filters often approximate curvature measures of numerically computed derivatives, i.e. gradient images, where meaningful features such as apses can be detected. Computing derivatives, however, comes with the drawback of smoothing an image, i.e. it overlooks details a human can detect. Furthermore, assumptions such as having 8 pixels as a neighbor to a central pixel or one height value per grid cell do not exist in 3D surface data.

Therefore, we choose to use numeric integration to prevent the smoothing effects of traditional derivative filters. Similarly to computing the area below a one-dimensional curve embedded in two-dimensional space, we compute the volume below our two-dimensional surfaces embedded in a three-dimensional space. This is achieved by considering each triangle of our 3D model as a top surface of a truncated prism extruded along an arbitrary axis, e.g. in z-direction, and as having arbitrary bottom surfaces, e.g. defined by the xy-plane. The sum off all volumes of all such prisms is the volume enclosed by the surface, that is, of our object acquired by a 3D scanner.

The moment we start computing subsets of the prism volumes, we start computing local curvatures. Choosing a sphere as the border of the subset makes the computed volume invariant against rotation. As we incorporate elements along the surface, the vertices of the triangular mesh become the centers of the spheres. The radius of the sphere is the parameter for the sensitivity of this volume integral invariant filter responding most perceptibly when the size of the feature is close to the radius of the sphere. To cover features of different sizes and multiple scales, we compute volume subsets for different radii. Therefore, this method is called *Multi-Scale Integral Invariant* (MSII) filtering (Mara 2012). Figure 1 shows two sets of concentric spheres for five different scales of a medieval seal. The integrated volume lies between 49% to

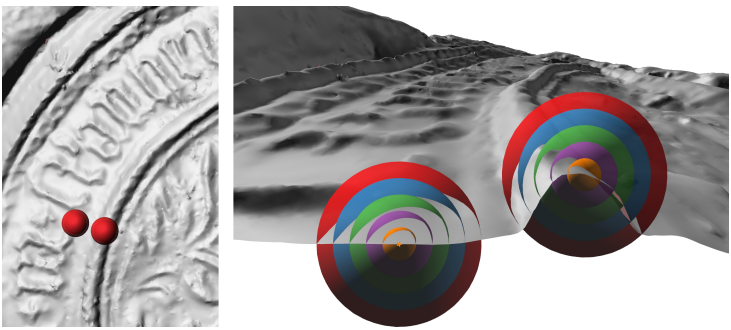


Figure 1: Triangular mesh describing the surface of a medieval seal in gray color (left). Detailed view shows two sets of concentric spheres used for local volume integration, where the volume below the gray surface and the sphere is computed (right).

51% for each sphere of the left set as each sphere is approximately cut in two halves by the surface. The integrated volume for the other set ranges from 40% of the volume of the smallest sphere to 22% of the volume for the largest sphere.

These ratios are a so-called *feature vectors* (or *functions*), which span a multi-dimensional feature space. Within this space, we can compute distance measures to a specific reference object, e.g. selected by pin-pointing a feature with a mouse click in a Graphical User Interface (GUI). Typical measures are the Euclidean distance or the Manhattan distance, but additional measures such as cross- or auto-correlation can also be applied. Considering the wide range of objects we encountered, there is always one suitable distance for each type of object. Finally, the distance measure is mapped to a color ramp, leading to a high contrast rendering of an object in false colors. Figure 2 shows a comparison of a photograph of a medieval seal and a 3D visualisation using a color ramp based on the colors of the Morgenstemming (Geissbuehler and Lasser 2013) which is suitable for printing in gray-scale and for colorblind persons.

Having determined features such as *Characters in 3D*, the next step is the feature extraction as a digital line drawing which can be made searchable by an approach based on a machine-learning technique known as *word spotting* (Kołcz et al. 2000). We illustrate this approach on one of the largest and oldest text sources known as cuneiform tablets.

2 From script in 3d to searchable line-drawings

For more than three millenia, scribes wrote documents using cuneiform script in the ancient Middle East (Soden 1994). Characters were typically written on clay tablets by imprinting a rectangular stylus and leaving a wedge (lat. *cuneus*) shaped trace, i.e.

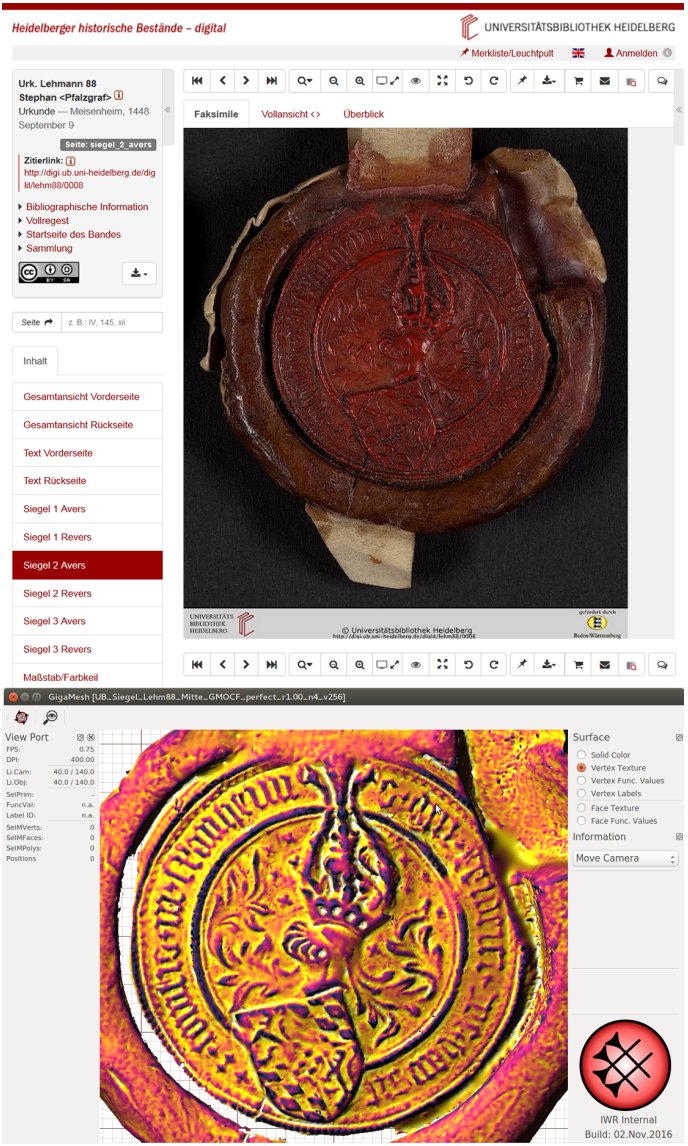


Figure 2: Comparison of a photograph of a seal and a visualization of its 3D measurement data. This is the 2nd seal of the document Lehmann 88, 9. September 1448, Meisenheim am Glan.

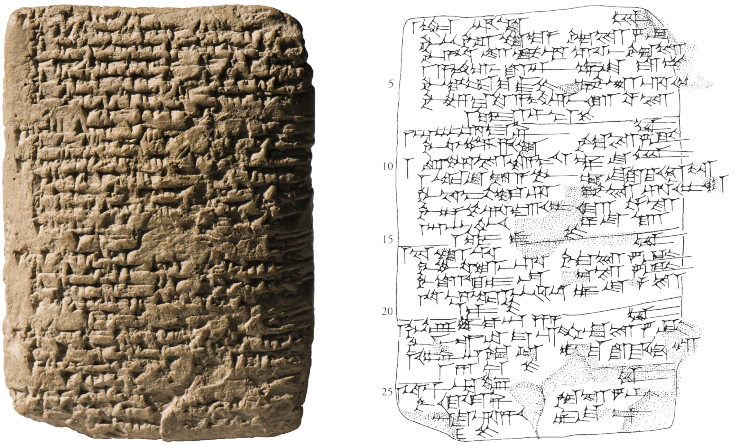


Figure 3: A cuneiform tablet and its tracing.

triangular markings, as shown on the left-hand side in figure 2. As clay was always cheaply and easily available, those capable of writing could produce a multitude of documents. Therefore, the content of cuneiform tablets ranges from mundane shopping lists to treaties between empires. There are hundreds of thousands of clay tablets preserved until this day thanks to their comparatively robust nature. In total, the amount of texts written in cuneiform script is comparable to those written in Latin or Ancient Greek. Important documents are, for example, the epic of *Gilgamesh* (Maul 2014), the declaration of the *Cyrus Cylinder* or the *Rosetta Stone*.

The increased availability of 3D representations and the tremendous amounts of 2D raster images of documents demand reliable methods for automated processing to keep tedious tasks such as drawing a cuneiform tablet to a minimum. This leads to the development of our *GigaMesh* software framework, which provides high-contrast images of 3D models (short for 3D measurement data) for improved readability of script in 3D. The *GigaMesh* framework was tested on numerous clay tablets with cuneiform script. The visualizations are achieved with the novel *Multi-Scale Integral Invariant* (MSII) filtering algorithm, applicable on the irregular triangular meshes describing a surface in 3D (Mara et al. 2010).

In a second step, our software framework was expanded by a line-tracing algorithm to extract features such as characters as *Scalable Vector Graphics* (SVG) (Mara and Krömker 2013), which describe the shape of extracted elements using the *eXtensible Markup Language* (XML). These two initial steps were also adapted for 2D raster images using *Dual Integral Invariant* filtering and the *potrace* algorithm to homogenize 3D and 2D sources (Bogacz et al. 2015a). The latter, in this case, are retro-digitized

manual drawings of cuneiform tablets. As experts today often use vector drawing tools such as *Inkscape* or *Computer Aided Design* (CAD) software, we have a third digital source, which can easily be exported in the SVG format. This makes digital manual drawings compatible with the automated drawings computed from 3D models using *GigaMesh*.

Having three homogenized digital data sources, the consequent step is the processing of SVG files to find repetitive patterns, e.g. groups of wedges of cuneiform script or any other graphical representation consisting of sets of prototypical elements. This work is done by application of *machine-learning* (ML) methods inspired by the idea of word spotting well known from the domain of *Handwritten Text Recognition* (HTR). This enables us to query a database of SVG files by using a drawing of a search word, character, or any other graphical element. In addition to the search capability, this approach enables future applications such as (i) automated annotation of characters, which is (ii) not limited to any writing system and can be used in other domains such as iconography or heraldry. Results are shown for synthetic data and real world data from more than six years of interdisciplinary projects at the interface between Applied Computer Sciences and the Humanities.

The *Cuneiform Digital Library Initiative* (CDLI) incorporates a number of projects aimed at cataloging cuneiform documents and making them available online as tracings, 2D images and sometimes as transliterations. However, none of these documents are annotated. Transliterations and translations are shown side-by-side with the photographs and retro-digitized scans of cuneiform tablets. For uninitiated readers it is impossible to correctly match the translated symbols to the respective symbols on the document.

Annotating documents manually is an arduous task that can only be performed by experts. The approach presented here can reduce the workload of annotating documents by repeating annotations on similar symbols automatically. This is accomplished by *symbol-spotting*, a concept similar to *word spotting* but extended to include graphical symbols. Symbols similar to those already annotated are spotted in a database and the respective annotations are applied. This approach reduces the workload to annotate documents. Each annotation is applied to the entire group of similar symbols, each time significantly reducing the symbols to be annotated. Figure 3 shows an annotated tablet with similar repeating symbols.

The unification of data sources requires a shared conceptual model of cuneiform wedges. The simplest possible description still allowing distinctive characters is a triangle representing the wedge-head with three associated arms representing the wedge arms. We use this description as our common shared model.

In born-digital tablets, our input data is a set of spline paths which we call strokes. These strokes are expressed as XML entities in the SVG source data. Wedges consist of up to six strokes, three for the triangular wedge-head and three for the wedge-arms.

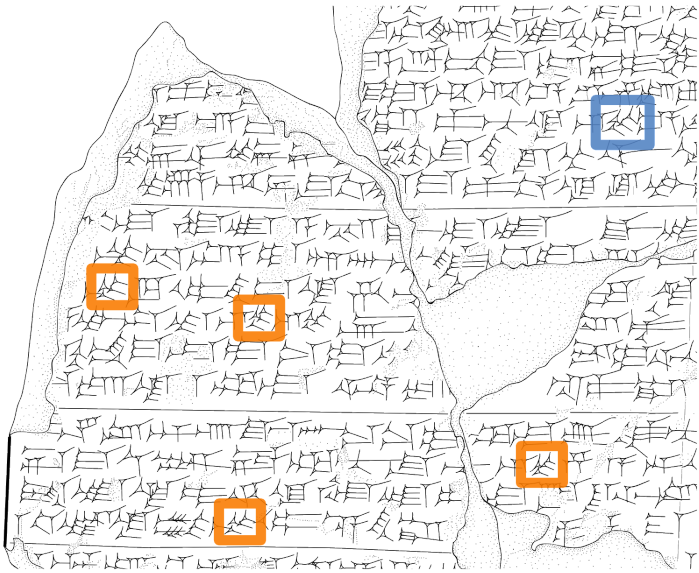


Figure 4: Annotation of a symbol (upper right corner) and automated repeated annotation of similar symbols (other markers).

We detect wedge-heads by finding three strokes intersecting pairwise. Wedge-arms are any additional strokes that intersect any stroke of the wedge-head.

This description of wedges is general enough to match all wedges on born-digital cuneiform transcriptions. It also matches many more structures which are not proper wedges, as can be seen in figure 4. One difficulty is that cuneiform script is written very densely. Strokes from different wedges may intersect and create false positives when analysing wedge heads or arms.

We meet this challenge by assuming that most strokes have been drawn to indicate proper wedges. We assign strokes to detected possible wedges. Strokes cannot fill two roles at once. Either (i) a stroke is assigned to be one of the three sides of a wedge-head or (ii) it is assigned to be one of the three wedge-arms. Strokes can also be left unused if drawn by error on a transcription. This task can be expressed as an optimal assignment problem, facilitating a computationally efficient solution.

Subsequent steps in our workflow and in a typical machine-learning workflow require a fixed size feature representing cuneiform characters. We model wedges using keypoints deriving directly from the way wedges are drawn in transcriptions. The

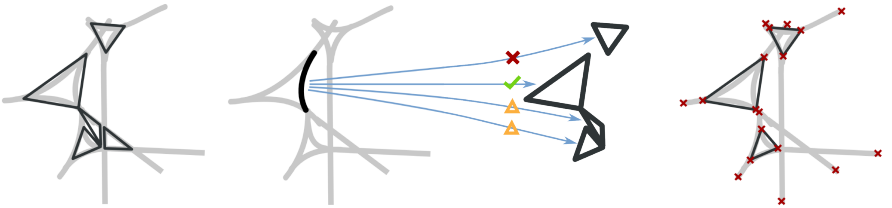


Figure 5: Wedge hypotheses (left), assignment of strokes to hypothesized triangles (mid) and final keypoint model of accepted modeled wedges (right).

keypoint feature-vector models wedges using six two-dimensional points, as shown in figure 4. The first three points are the vertices of the three strokes intersecting pairwise, forming the wedge-head. The last three points are endpoints of the wedge-arms attached to the respective wedge vertices. This model is described in detail in Bogacz et al. (2015b). We also successfully utilized this representation in a machine-learning workflow to extract repeated cuneiform patterns (Bogacz and Mara 2016a).

Part-structured models provide means to describe geometrical objects by the relationships of their components. Howe (2015) has presented a part-structured model based on point centers and a tree of flexible, spring-like links inbetween. Additionally, he also introduced highly efficient means of these models' parallel computation. We adapt and model cuneiform symbols using a part-structured model of wedges connected by spring-like links. The generalized distance transform (GDT) employed by Howe is modified to use the Euclidean distance between the keypoint feature-vectors of wedges. Symbol-spotting is then performed by transforming the document regarding the query and computing the distance field (Bogacz et al. 2016b). Then, local minima are possible locations of the query symbol in the document. Figure 5 shows exemplary key stages of this process.

We evaluated our methods on a dataset of two cuneiform tablets line-traced by assyriologists. A vector graphics editor has been used to create born-digital SVG files. Each of these tablets contain approximately 500 identifiable cuneiform characters on each side.

We performed retrieval queries by example, using the set of segmented cuneiform characters. For each result returned, an expert decided whether it belonged to the class of the query and tagged it with either true positive or false positive. Additionally, we evaluated our method against the work of Rothacker et. al. (2013) on word spotting on Latin script. Their work on cuneiform detection (Rothacker et al. 2015) could not be evaluated since elevation data, as used in their approach, was not available

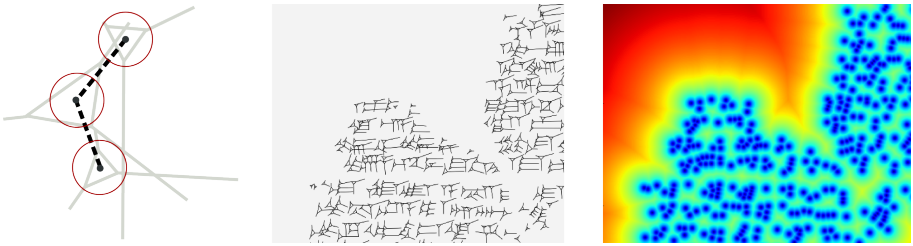


Figure 6: Balls and flexible springs model (left), keypoints of the document to be searched (middle) and resulting distance field after transformation (right).

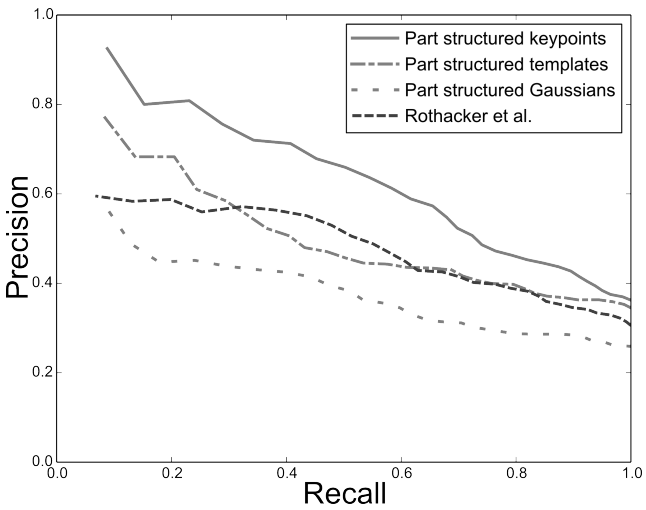


Figure 7: Precision and recall of the symbol-spotting approach presented here. We compare our different wedge models and the work by Rothacker et al. on word spotting.

for our dataset. In general, there currently is no standardized dataset of cuneiform tablets for learning tasks available by means of Open Access in a manner the George Washington letters are for Latin word spotting. Figure 6 shows the precision-recall plot of our three part structured algorithms including the approach as suggested by Rothacker et. al. (2015).

In addition to the keypoint model, we also experimented with other wedge models and evaluated these on our dataset. The native keypoint model presented the best performance and outperformed the state of the art in Latin word spotting significantly. Our approach has been modeled to exploit the geometrical properties of cuneiform

and works on vector data instead of raster data. Therefore, the search word for a query is actually a drawing, which leads to an automatable annotation by querying all possible cuneiform signs as found in symbol lists (Borger 2010).

3 Conclusion and outlook

Even in the case of cuneiform tablets belonging to the oldest important text sources existing in vast numbers, there were virtually no computational methods available to assist crucial tasks like examination or transcription when we began developing digital tools for the analysis of clay tablets in 2009. In the first phase, we established a workflow for digitization of the clay tablets using optical metrology resulting in high-resolution 3D models. Afterwards, we developed a robust algorithm using Multi-Scale Integral Invariant filtering for high-contrast visualizations of *Script in 3D*, which was implemented as modular GigaMesh software framework. Due to its versatile nature forgoing any inclusion of a-priori knowledge and complex parametrization, we could successfully apply the framework to e.g. Roman inscriptions and weathered medieval Jewish epitaphs. For the latter, we could recover approximately 20% of additional characters which had been declared to be lost forever.

The second phase outlined in this article utilizes digital drawings of the cuneiform computed from 3D measurement data. These drawings are XML-based Scalable Vector Graphics and act as an interface to manual drawings which can be incorporated by homogenization for both retro-digitized and born-digital data. Using a minimalistic geometric model (template) to describe the wedges, i.e. radical element of cuneiform script, we were able to establish search capabilities based on word spotting. Our search algorithm enables the user to query by drawing instead of query by some sort of encoded symbol. Therefore, we can treat any cuneiform writing independent of any underlying language. This is a key factor as there are several major languages originating from at least three different language families sharing cuneiform script. Together with diverse local dialects and challenges like the *UD.GAL.NUN* signs (Zand 2016), it appears that techniques commonly applied in Computational Linguistics are prone to become isolated applications.

The whole processing workflow from high-resolution 3D measurement data to searchable drawings contains many modules to be reused for other projects within the Digital Humanities. Filtering techniques are adaptable to the domain of raster images provided by photographs and flatbed scanners. Examples are the anistropic filtering of rubbings of ancient Sutra chiseled into stone walls (Mara et al. 2009), material structures, i.e. unique stripe patterns of Papyri (Mara and Sänger 2013), or the improvement and vectorization of faded George Washington letters (Mara 2016).

The latter will be a future challenge to adopt the symbol-spotting of cuneiform – which is actually a handwriting in 3D – to handwriting with pen and paper in 2D.

Bibliography

Bertrand, Joseph, C.F. Diquet, and Victor Puiseux. "Démonstration d'un théorème de Gauss." *Journal de Mathématiques* 13 (1848). 80–90.

Bogacz, Bartosz and Hubert Mara. "Clustering Fundamental Spatial n-Grams for Large Scale Cuneiform Search." *Proceedings of the International Conference on Document Analysis Systems (DAS)* 2016.

Bogacz, Bartosz, Judith Massa, and Hubert Mara. "Homogenization of 2D & 3D Document Formats for Cuneiform Script Analysis." *Proceedings of the 3rd Conference on Historical Document Imaging and Processing (HIP)* 2015.

Bogacz, Bartosz, Nicholas Howe, and Hubert Mara. "Segmentation Free Spotting of Cuneiform using Part Structured Models." *International Conference on Frontiers in Handwriting Recognition (ICFHR)* 2016.

Bogacz, Bartosz, Michael Gertz, and Hubert Mara. "Character retrieval of vectorized cuneiform script." *International Conference on Document Analysis and Recognition* 2015.

Borger, Rykle. "Mesopotamisches Zeichenlexikon." *Alter Orient und Altes Testament – Veröffentlichungen zur Kultur und Geschichte des Alten Orients und des Alten Testaments (AOAT).* Volume 305, 2nd edition. Münster: Ugarit-Verlag, 2010.

Boss, Martin, Bernd Breuckmann, and Matthias Göbbels. "Auf der Spurensuche des Handwerks zum Prägen antiker Münzen unter Einsatz von höchstaufgelösten digitalen 2D- und 3D-Modellen." *Konferenzband EVA 2012 Berlin Elektronische Medien & Kunst, Kultur, Historie.* 2012. 73-78.

CDLI: *Cuneiform Digital Library Initiative.* <http://cdli.ucla.edu/>.

Effinger, Maria, Eberhard Pietzsch, and Ulrike Spyra. "Digitalisierung und Erschließung spätmittelalterlicher Bilderhandschriften aus der Bibliotheca Palatina ein Kooperationsprojekt der Universitätsbibliothek und des Kunsthistorischen Instituts der Universität Heidelberg." In Thaller, Manfred (ed.). *Fundus – Forum für Geschichte und ihre Quellen* 5 (2003). 62-89.

Gauss, Carl Friedrich, James Caddall Morehead, and Adam Miller Hiltebeitel. *General Investigations of Curved Surfaces of 1827 and 1825.* Eberdeen: Watchmaker Publishing, 2007.

Geissbuehler, Matthias, and Theo Lasser. "How to display data by color schemes compatible with red-green color perception deficiencies." *Optics express* 21.8 (2013). 9862-9874.

Howe, Nicholas R. "Inkball models for character localization and out-of-vocabulary word spotting." *Proceedings of the International Conference on Document Analysis and Recognition* 2015.

Kołcz, Aleksander. et al. "A line-oriented approach to word spotting in hand-written documents." *Pattern Analysis & Applications* 3.2 (2000). 153–168.

Krömker, Susanne. "Neue Methoden zur besseren Lesbarkeit mittelalterlicher Grabsteine am Beispiel des Heiligen Sands in Worms." *Die SchUM-Gemeinden Speyer - Worms - Mainz. Auf dem Weg zum Welterbe.* Regensburg: Schnell & Steiner, 2013. 167–188.

Mara, Hubert, Jan Hering, and Susanne Krömker. "GPU based optical character transcription for ancient inscription recognition." *Proceedings of the 15th International Conference on Virtual Systems and Multimedia (VSMM)* 2009.

Mara, Hubert et al. "GigaMesh and Gilgamesh - 3D Multi Scale Integral Invariant Cuneiform Character Extraction." *Proceedings of the International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST)* 2010. 131–138.

Mara, Hubert and Susanne Krömker. "Vectorization of 3D-Characters by Integral Invariant Filtering of High-Resolution Triangular Meshes." *Proceedings International Conference on Document Analysis and Recognition (ICDAR)* 2013. 62-66.

Mara, Hubert, and Julia Portl. "Acquisition and Documentation of Vessels using High-Resolution 3D-Scanners." In Trinkl, E. (ed.). *Neue interdisziplinäre Dokumentations- und Visualisierungsmethoden, Corpus Vasorum Antiquorum Österreich, Beiheft 1.* Wien: Verlag der Österreichischen Akademie der Wissenschaften, 2013. 25–40.

Mara, Hubert, and Patrick Sänger. "Präzise Bestimmung von Materialstrukturen bei Papyri mit 3D-Messtechnik." *Zeitschrift für Papyrologie und Epigraphik (ZPE)* 185 (2013). 195–199.

Mara, Hubert. "Multi-Scale Integral Invariants for Robust Character Extraction from Irregular Polygon Mesh Data." PhD thesis at Heidelberg University, 2012. Online/Open Access. URN: urn:nbn:de:bsz:16-heidok-138909.

Mara, Hubert. "Made in the humanities: Dual integral invariants for efficient edge detection." *it–Information Technology* 58.2 (2016). 89–96.

Maul, Stefan M. "Das Gilgamesch-Epos". 6th edition. Munich: C.H.Beck, 2014.

Roberts, Lawrence G. "Machine Perception Of Three-Dimensional Solids." Massachusetts Institute of Technology, Lincoln Laboratory, 1963.

Rothacker, Leonard et al. "Retrieving cuneiform structures in a segmentation-free word spotting framework." *Workshop on Historical Document Imaging and Processing* 2015.

Rothacker, Leonard, Marcal Rusinol, and Gernot A. Fink. "Bag-of-features HMMs for segmentation-free word spotting in handwritten documents." *International Conference on Document Analysis and Recognition* 2013.

Sablatnig, Robert, and Christian Menard. "Stereo and Structured Light as Acquisition Methods in the Field of Archaeology". *Mustererkennung* (1992). 398–404.

Soden, Wolfram von. "The ancient Orient: an introduction to the study of the ancient Near East." *Wm. B. Eerdmans Publishing Co.,* 1994.

Ullman, Shimon. "The Interpretation of Structure from Motion." *Proceedings of the Royal Society B* 203.1153 (1979). 405–426.

Willems, Geert et al. "Easy and cost-effective cuneiform digitizing." *6th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST*, 2005.

Woodham, Robert J. "Photometric method for determining surface orientation from multiple images." *Optical Engineering* 19.1 (1980). 139-144.

Zand, Kamran Vincent. *"UD.GAL.NUN".* Reallexikon der Assyriologie und Vorderasiatischen Archäologie 14, 2016. 271-273.

Automatic Dating of Historical Documents

Vincent Christlein, Martin Gropp, Andreas Maier

Abstract

With the growing number of digitized documents available to researchers it is becoming possible to answer scientific questions by simply analyzing the image content. In this article, a new approach for the automatic dating of historical documents is proposed. It is based on an approach only recently proposed for scribe identification. It uses local RootSIFT descriptors which are encoded using VLAD. The method is evaluated using a dataset consisting of context areas of medieval papal charters covering around 150 years from 1049 to 1198 AD. Experimental results show very promising mean absolute errors of about 17 years.

Zusammenfassung

Mit der steigenden Zahl der für Forscher zugänglichen digitalisierten Dokumente wird es möglich, wissenschaftliche Fragestellungen durch die einfache Analyse der Bilddaten zu beantworten. In diesem Beitrag wird ein neues Vorgehen für die automatische Datierung historischer Dokumente vorgestellt. Es basiert auf einem Ansatz, der erst vor kurzem für die Schreiberidentifikation entwickelt wurde und nutzt lokale RootSIFT-Deskriptoren, die mit VLAD codiert werden. Die Methode wird mit einem Datensatz evaluiert, der aus den Textbereichen mittelalterlicher Papsturkunden aus rund 150 Jahren (1049-1198) besteht. Experimentelle Ergebnisse zeigen eine sehr vielversprechende mittlere Fehlerrate von rund 17 Jahren.

1 Introduction

Dating historical documents can be a time-consuming and expensive process which typically requires the consultation of experts of history and/or paleography. While the chemical analysis of the paper through radiocarbon dating often yields reasonable accuracy, at least for the time of production of the writing medium, non-invasive methods are often preferable for a variety of reasons.

These approaches can be divided into two groups: content-based methods and image-based methods. Content-based methods relate to procedures which derive the date of production from information in the text. Either directly, e. g., an event directly referred to in the text can be related to a known date. Or indirectly, through

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 151–164.

a linguistic analysis of the text, see for example the work of Feuerverger et al. (2008), who dated manuscripts from the 11th to the 15th century. This is possible if enough dated reference material exists.

This is also a prerequisite for image-based methods. In contrast to content-based methods, however, the text does not need to be transcribed first. For several manuscripts, a rough date can be estimated (manually or automatically) based on the layout of the document or the symbols/images it contains. In papal charters, for example, there typically is a *rota* symbol containing the name of the pontificate. Moreover, the handwriting can give a clue to the date since different handwriting styles were used in different periods of time . By extracting these information, a semi or fully automatic program can assist a paleographer in dating handwritten documents. It is also to be noted that large-scale dating, i. e., the dating of hundreds of manuscripts or more, might be too time-consuming for an individual. Here, automated methods suggesting a probable date might be useful for initial estimates or may also point out interesting documents to the researcher. For example, outliers in a large corpus of documents might just relate to an interesting handwriting — or the style could actually point towards a later date than the content, indicating a potential document forgery.

Wahlberg et al. (2016) showed that content- and image-based methods can also be combined for an improved automatic dating.

Automatic dating may also help to improve OCR quality as specialized classifiers can be trained for specific date ranges when they are known. Li et al. (2015) have shown great improvements in OCR when estimating the date of printed manuscripts in advance.

Algorithmically, the dating of handwritten text is closely related to the problem of (automatic) writer identification.[1] But while there are fixed classes of writers in the case of writer identification, image-based dating is typically seen as having a regression problem, i. e., we determine continuous targets (the dates) instead of fixed classes (the writers).

In this paper, we propose a new method for automatic dating. The individual parts of the approach have already been used successfully for writer identification (Christlein et al. 2014; Christlein et al. 2015). These publications draw on clean benchmark datasets, while this work relies on experiments with historical documents. Historical documents are typically digitized in high definition. Thus, we evaluate different strategies to lower the computational burden. Moreover, historical documents often contain large deficiencies such as holes or stains. We evaluate different strategies for feature sampling and study their effects on dating accuracy. An example image can be seen in figure 1.

[1] Note: "writer" and "scribe" are used interchangeably throughout this paper.

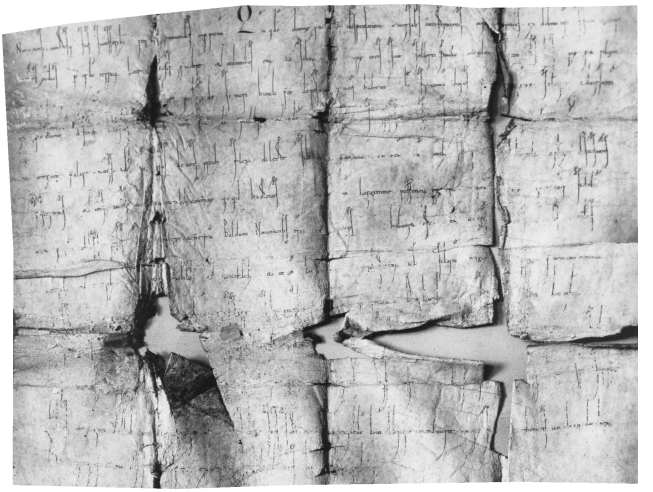


Figure 1: Image excerpt of a papal charter. Jaffé / Loewenfeld no. 4671; pontificate: Alexander II; date: January 28, 1070; image courtesy of the Göttingen Academy of Sciences and Humanities.

This paper is organized as follows: after the related work is presented in section 2, our proposed method is explained in section 3. Section 4 covers the evaluation of our experiments and results. The paper is concluded in section 5.

2 Related work

Dating of historical manuscripts

Automatic image-based dating of historical manuscripts is a relatively new discipline with virtually no visible research until only a few years ago, which was probably owed to the lack of sufficiently large digitized collections of suitable documents. In 2014, He et al. presented a new dataset and used Hinge and Fraglets features in a nested SVR approach to predict the year of a document's creation. In the following year, Wahlberg et al. (2015) proposed a method focused in particular on low-quality images, based on shape context and Stroke Width Transformation. Recently, He et al. (2016a) added a new unsupervised attribute learning step and finally treated document dating as a classification problem, an approach they continue in their later work (He et al. 2016b) with local contour fragments and stroke fragments features. While Wahlberg et al. advance to place special emphasis on incorporating language information in their 2016 paper, requiring manual transcriptions that are not easily available in many cases, they also continue to improve their solely image-based method.

Handwriting classification

The problem of dating handwritten text is methodically similar to text style recognition or writer identification. Writer identification can be categorized into two groups: *textural* methods and *allograph*-based methods. In textural based methods, comprehensive statistic information are computed from the handwriting, e. g., the width of the ink stroke. A prominent example describes the handwriting by means of local binary patterns (Nicolaou et al. 2015). In comparison, allograph-based methods rely on a background model computed from local descriptors of a training set. This background model is then used to *encode* the local descriptors, i. e., to compute statistics from them. The most closely related publications belong to the latter group (Christlein et al. 2014; 2015a; 2015b). In our earliest work (Christlein et al. 2014), we used RootSIFT descriptors as local descriptors in combination with GMM supervectors for encoding. A variant of the GMM supervectors was also used in our most recent work (Christlein et al. 2015b), where they are employed to encode CNN activation features. In contrast, *vectors of locally aggregated descriptors* (VLAD) are used to encode Zernike moments which were evaluated densely at the script contour in our other work (Christlein et al. 2015a). This approach won the ICDAR 2015 competition on multi-script writer identification (Djeddi et al. 2015).

3 Methodology

Since the contour extraction involves more steps in historical documents than for clean benchmark data, we employ sparsely sampled RootSIFT descriptors (Arandjelović and Zisserman 2012) for our baseline method. For the aggregation of these local descriptors, we use multiple VLAD encodings (Jégou and Chum 2012; Jégou et al. 2012). The global descriptors of the training set are used to train a classifier for the date prediction.

The full workflow consists of three main steps: 1) local feature extraction, where we employ RootSIFT descriptors, 2) the aggregation of the local feature descriptors in the encoding step, 3) estimation of the date by means of linear regression.

3.1 Feature extraction

We make use of the Scale Invariant Feature Transform (SIFT) (Lowe 2004). SIFT descriptors encode the orientations of gradients in the neighborhood of scale and rotation invariant positions (*keypoints*) in the image. Note that we set the keypoint-angles to zero, since rotation-invariance is not necessary for the classification of handwriting (Fiel and Sablatnig 2013). Each SIFT descriptor is normalized using the Hellinger kernel (Arandjelović and Zisserman 2012), i. e., the square root is applied to each element, hence the name *RootSIFT*. This normalization reduces the effect of

dominating values in the SIFT descriptor and has been shown to be very beneficial for writer identification (Christlein et al. 2014).

3.2 Encoding

The formation of a global descriptor is accomplished by the use of VLAD (Jégou et al. 2012). First, a dictionary C is computed from local descriptors using k-means. It consists of K cluster centers $\mu_k \in \mathbb{R}^D,\ k \in \{1, \dots, K\}$. For each cluster, all residuals between the cluster center and its nearest local descriptors are aggregated. Formally, given T as local descriptors $\mathrm{x_t} \in \mathbb{R}^D,\ t \in \{1, \dots, T\}$ of a single image:

$$\mathrm{v}_k = \sum_{\mathrm{x}_t:\ \mathrm{NN}(\mathrm{x}_t)=\mu_k} (\mathrm{x}_t - \mu_k)\,, \tag{1}$$

where NN(x) denotes the nearest neighbor of x. Then, the full $K \times D$ dimensional global descriptor follows by concatenation:

$$\mathrm{v} = (\mathrm{v}_1^\top, \dots, \mathrm{v}_K^\top)^\top\,. \tag{2}$$

Jégou and Chum (2012) showed that it is beneficial to use more than one dictionary resulting in multiple global descriptors. These are jointly decorrelated and dimensionality is reduced by means of PCA whitening. This has also been shown to improve the results for writer identification (Christlein et al. 2015a).

3.3 Date regression

To estimate the date the decorrelated VLAD vector v is used in a linear Support Vector Regression (SVR). The best hyper-parameters for the SVR are selected in an inner 5-fold cross-validation.

4 Evaluation

In this section, we introduce the dataset and error metrics that we use for evaluation. For the evaluation, we conduct several experiments using different preprocessing and sampling techniques for the feature extraction.

4.1 Dataset

The dataset used for evaluating the date estimation consists of 697 digitized medieval papal charters with known date. The documents come from three different archives. The majority (580) were provided by the Göttingen Academy of Sciences and Humanities (papsturkunden.de), 67 charters are provided by the Collaborative Archive

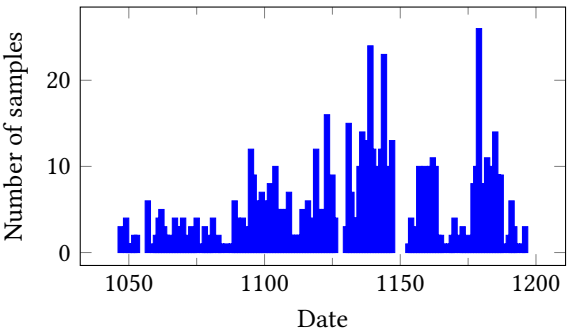


Figure 2: Distribution of documents in the dataset over the years.

Monasterium.net (Mom), and 50 stem from the Lichtbildarchiv älterer Originalurkunden - Philipps Universität Marburg (LBA). Most digitized images are retro-digitizations, i. e., digitizations from analog photos. Thus, the resolution and size of the documents vary greatly. Many documents also contain characteristics such as folds, stains, and rips, cf. figure 1. Also note that documents from the LBA contain a watermark which might have a small effect on the test accuracy (although the test set only contains two LBA-charters). The charters consist of one single document image. As a consequence, our experiments are inherently document-independent. We cannot guarantee an evaluation independent of the writer because the scribal hand is not known for the majority of the corpus. For in the time between 753 and 1197 AD, around 25 000 papal charters are handed down, about 20 000 of which are dated to the 12$^{\text{th}}$ century (see Hiestand 1999, 4), the chance of finding the same scribal hand in two different charters is presumably quite low. The dates of the charters of our dataset range between 1047 and 1196 AD. The year-sample distribution is depicted in figure 2.

We do not use the complete charters, but only the main context area, see figure 1 for an example. This way, it is guaranteed that graphical symbols (*rota*, *benevalete*, etc.) do not influence the results, and only the handwriting style is used for the date estimation. The main context areas were annotated during the project *Script and Signs. A computer-based analysis of high medieval papal charters. A key to Europe's cultural history* (PuhMa). We randomly split the dataset in roughly independent training (630 documents) and test (69 documents) sets.

4.2 Error metrics

We evaluate the predicted years of writing according to several error metrics. The *Mean Absolute Error* score (MAE) provides a simple indication of the average performance of the estimator:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right| , \tag{3}$$

Variant	MAE	RMSE
Baseline	20.62	25.09
Baseline w.o. extr. kpts	20.81	25.21

Table 1: Evaluating the influence of extremely sized keypoints. The first row shows the results for the baseline, while the second row shows the results for the baseline method without extremely sized keypoints.

where N is the number of test documents, and y_i and $\hat{y}_i$ are the true and estimated years for document i, respectively.

In order to gain some more insight into the behavior regarding outliers, we consider another metric, the Root Mean Squared Error (RMSE), which puts more emphasis on outliers than MAE:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}. \tag{4}$$

Finally, the *Cumulative Score* (CS) (Geng et al. 2007) is a useful metric in cases where there is no or little value in a perfectly exact prediction. Instead, it assumes an *acceptable error* α (here given in years) and gives the percentage of the predictions that fall within this margin of error:

$$\text{CS}_\alpha = \frac{N_{e\leq\alpha}}{N} \cdot 100\%. \tag{5}$$

4.3 Experiments

We evaluate different aspects regarding the size and sampling strategies for the RootSIFT descriptors. First, we try to limit the number of descriptors, next we experiment with reducing the image size. Finally, we evaluate the impact of different sampling strategies.

The *baseline* in our experiments denotes the pipeline as explained in section 3. Table 1 shows that the baseline approach gives an MAE of about 20 years and an RMSE of 25 years. According to the literature (see section 2), this is comparable to the state of the art in image-based dating. It follows that the transfer from a writer identification method to a date estimation method was successful.

Reducing the number of descriptors

In a first experiment, we removed keypoints varying more than twice the standard-deviation from the mean keypoint size. This way, non-standard keypoints are removed.

Variant	MAE	RMSE
Baseline (unscaled)	20.62	25.09
Scale 2048	21.97	26.53
Scale 1024	40.67	47.28
Center-crop 2048	23.52	28.65
Center-crop 1024	32.56	38.84

Table 2: Comparison of the unscaled baseline results with scaled, or cropped versions of the image.

More formally, a keypoint k is removed when:

$$s(k) \gtrless \mu \pm 2 \cdot \sigma\,, \tag{6}$$

where $s(k)$ is the size of the keypoint k, and μ and σ are the mean and standard deviation of all keypoint sizes in the image respectively. See for example figure 3b, where the orange keypoints denote the extreme keypoints, i. e., those omitted for this experiment. Interestingly, table 1 shows that this step is not advisable in comparison to the baseline raised by the RMSE and MAE. It seems that larger keypoints, which result in descriptors covering a larger image portion, are beneficial. Thus, we do not remove extremely sized keypoints in the following experiments.

Influence of image scaling

Next, we evaluate the impact of image scaling. Since the images are quite large (in average 2603 × 2021 pixels), this would decrease the computational load. Thus, we scale down the images in such a way that the larger dimension consists of 2048 (1024) pixels by retaining the aspect ratio of the image. In two subsequent experiments, we take the center-crop of 2048 × 2048 pixels (1024 × 1024). If one image dimension is smaller we take this dimension, i. e., $\min(2048, \text{width}) \times \min(2048, \text{height})$, proceeding similarly for center-crops with 1024 pixels in each dimension.

Table 2 shows that any scaling harms the date estimation. However, results for the unscaled baseline are only slightly better than a moderate scaling of 2048 pixels. A scaling to 1024 pixels worsens the results drastically. A possible reason might be the lower number of detected keypoints, and, thus, extracted descriptors in the image. Using the center-crop of 2048 pixels is slightly worse than rescaling to 2048 pixels. Interestingly, the center-crop of 1024 pixels is much better than its scaling counterpart.

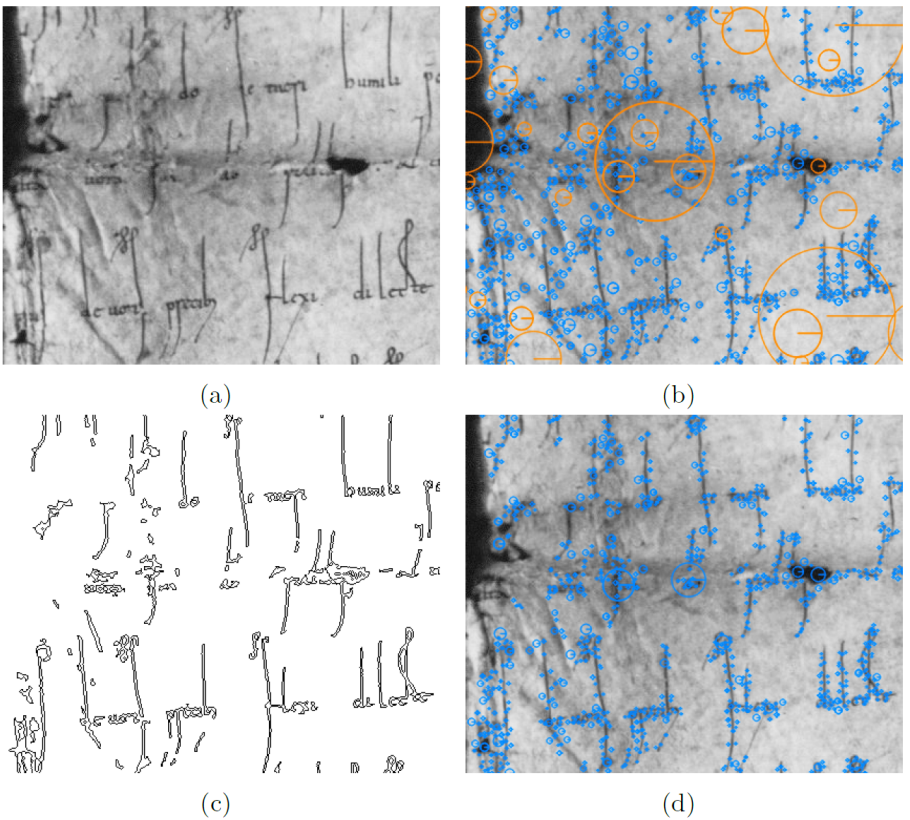


Figure 3: a) excerpt of figure 1; b) SIFT keypoints (orange: keypoints with extreme size), for the baseline results all keypoints are taken; c) contour sampling; d) masked SIFT keypoints.

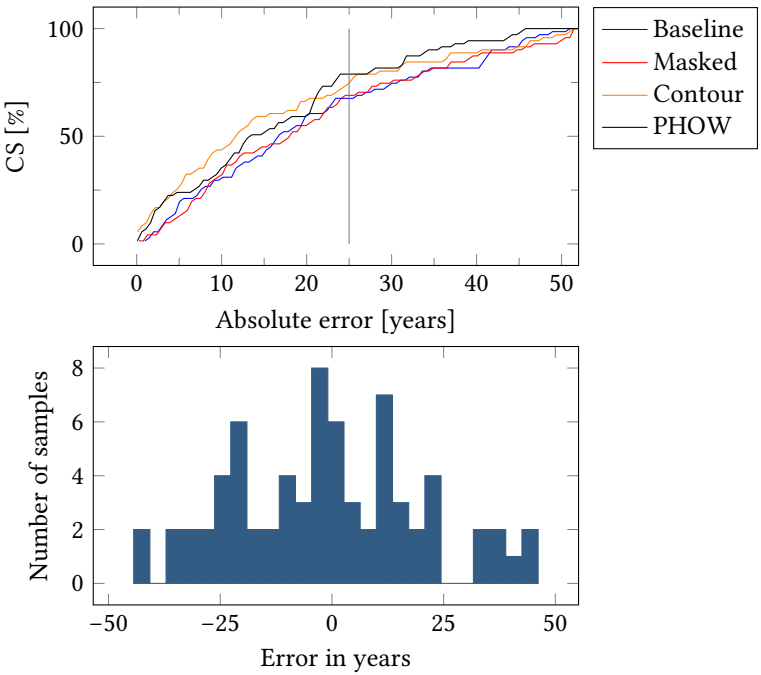


Figure 4: Top: Cumulative absolute error distributions of the different sampling strategies. Bottom: Histogram (with 25 bins) of errors using the PHOW method.

Influence of feature sampling

As a last experiment we evaluated different sampling strategies, i. e., we compare different positions (keypoints) at which feature descriptors are computed. The baseline uses the original SIFT keypoint detection proposed by Lowe (2004), see for example figure 3b. At these keypoints, the RootSIFT descriptors are extracted. We compare it with three different variants:

1.) We compute the keypoints as before but use only those which are close to the handwriting script (denoted as Masked RootSIFT). Therefore, we compute a mask which mainly consists of handwriting. To segment the handwriting in background and text, we apply the binarization technique proposed by Su et al. (2010). Remaining noise is reduced by removing connected components thare are too small (less than 20 points) or too large (larger than 3000 points). The mask is dilated by a 5×5 circular shape to allow keypoints at the border of the handwriting. Figure 3d shows an example of masked keypoints.

Variant	MAE	RMSE	CS_{25}
Baseline	20.62	25.09	67.61
Masked RootSIFT	21.45	27.19	69.01
Contour RootSIFT	17.17	23.01	73.24
PHOW	**16.95**	**21.04**	**78.87**

Table 3: Evaluation of different sampling strategies.

2.) We use the points situated at the contour of the handwriting (denoted as Contour RootSIFT). Therefore, we use the same strategy as before without the dilation step, see for example figure 3c for the extracted contour. At each contour point we evaluate the RootSIFT descriptor.
3.) Finally, a fast and dense variant of SIFT, known as Pyramid Histogram of Visual Words (PHOW) (Bosch et al. 2007) is computed. We extract the PHOW descriptor from the slightly downscaled version where the larger image dimension was resized to 2048 pixels. Descriptors having a norm lower than 0.05 were discarded since they stem from homogeneous areas. We used multiple bin sizes (4, 10, 16) and a step size of 10. The descriptors are Hellinger-normalized, similarly to the RootSIFT descriptors.

Table 3 shows the results for the four different strategies. It reveals that the masked variant of RootSIFT slightly worsens the results. This might be related to parts where the segmentation for the mask creation fails. In contrast, the contour-based RootSIFT and the densely sampled RootSIFT descriptors both surpass the baseline results by a significant margin. Both achieve similar results of about 17 years MAE. Regarding the RMSE, PHOW is in favor. Note, however, more keypoints are computed for these two methods and an order of magnitude more than for the baseline. This effects the computational costs for the feature extraction (especially for the contour-based method) and for the encoding step since more descriptors need to be accumulated.

The CS_{25} draws a picture similar to the MAE and RMSE values. However, figure 4 (top) shows that in ranges below 20 years, the cumulative score of the contour-based sampling strategy is in favor. Figure 4 (bottom) depicts the error histogram of the PHOW method. While there are fewer documents outside errors of ±25 years, there is a clear peak around 0 showing that several documents could be dated very exactly. Note that the results show a significant (Pearson-)correlation of 92% between the regression output and the true date (significance level 0.001).

5 Conclusion

In this work, we have shown that a method originally developed for writer identification can be transferred to fulfil the task of dating historical manuscripts. The historical manuscripts we used are not comparable to clean benchmark data, they are typically digitized in high resolution but contain deficiencies such as holes or stains. For this reason, we evaluated different strategies to lower the computational burden by reducing the image size. The results show that, while moderate scaling is acceptable, the results drop drastically in case of excessive scaling.

We also showed that sampling strategies other than SIFT keypoints improve the results. Both a dense SIFT variant (PHOW) as well as contour-based sampling surpass the baseline achieving an MAE of about 17 years and an RMSE of 21 years. However, the increase in keypoints comes at the cost of an increased computational complexity.

For future research, we would like to expand our studies regarding the feature sampling. Maybe other keypoint strategies, such as a sparse contour sampling could decrease the computational cost. Given enough training data, deep learning techniques could also be used for dating handwritten text similar to the work of Li et al. (2015).

Acknowledgments

Many thanks to Benedikt Hotz, Benjamin Schönfeld, Thorsten Schlauwitz, and Viktoria Trenkle for the annotation of the papal charters.

Bibliography

Arandjelović, Relja, and Andrew Zisserman. "Three things Everyone Should Know to Improve Object Retrieval." *Computer Vision and Pattern Recognition* (CVPR). IEEE Conference. 2012. 2911–2918.

Bosch, Anna, Andrew Zisserman, Xavier Mu, and Xavier Munoz. "Image Classification Using Random Forests and Ferns." *IEEE 11th International Conference onComputer Vision* (ICCV). 2007. 1–8.

Christlein, Vincent, David Bernecker, and Elli Angelopoulou. "Writer identification using VLAD encoded contour-Zernike moments." *Document Analysis and Recognition* (ICDAR). 13th International Conference. 2015. 906–910.

Christlein, Vincent, David Bernecker, Florian Hönig, and Elli Angelopoulou. "Writer Identification and Verification Using GMM Supervectors." *2014 IEEE Winter Conference on Applications of Computer Vision* (WACV). 2014. 998–1005.

Christlein, Vincent, David Bernecker, Andreas Maier, and Elli Angelopoulou. "Offline Writer Identification Using Convolutional Neural Network Activation Features." In Gall, Juergen,

Peter Gehler, and Bastian Leibe (eds.) *37th German Conference on Pattern Recognition.* GCPR 2015. Aachen, Germany, October 7-10, 2015. Proceedings. London: Springer International Publishing, 2015. 540–552.

Djeddi, Chawki, Somaya Al-Maadeed, Abdeljalil Gattal et al."ICDAR2015 Competition on Multi-script Writer Identification and Gender Classification using 'QUWI' Database." *13th International Conference on Document Analysis and Recognition* (ICDAR). 2015. 1191–1195.

Feuerverger, Andrey, Peter G. Hall, Gelila Tilahun, and Michael Gervers "Using Statistical Smoothing to Date Medieval Manuscripts." In *Beyond parametrics in interdisciplinary research: Festschrift in honour of Professor Pranab K. Sen.* 2008. 321–331. <http://arxiv.org/abs/0805.2490>.

Fiel, Stefan and Robert Sablatnig. "Writer Identification and Writer Retrieval using the Fisher Vector on Visual Vocabularies." *12th International Conference on Document Analysis and Recognition* (ICDAR). 2013. 545–549.

Geng, Xin, Zhi-Hua Zhou, and Kate Smith-Miles. "Automatic Age Estimation Based on Facial Aging Patterns." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 29.12 (2007). 2234–2240.

Gervers, Michael. "The DEEDS project and the development of a computerised methodology for dating undated English private charters of the twelfth and thirteenth centuries." *Dating Undated Medieval Charters.* 2000. 13–35.

He, Sheng, Petros Samara, Jan Burgers, and Lambert Schomaker. [2014.] "Towards Style-Based Dating of Historical Documents." *International Conference on Frontiers in Handwriting Recognition* (ICFHR). 2014. 265–270.

He, Sheng, Petros Samara, Jan Burgers, and Lambert Schomaker. [2016a.] "Historical Document Dating using Unsupervised Attribute Learning." *12th IAPR Workshop on Document Analysis Systems* (DAS). 2016. 36–41.

He, Sheng, Petros Samara, Jan Burgers, and Lambert Schomaker. [2016b.]"Image-based Historical Manuscript Dating Using Contour and Stroke Fragments." *Pattern Recognition 58.* 2016. 159-171. DOI: 10.1016/j.patcog.2016.03.032.

Hiestand, Rudolf. "Die Leistungsfähigkeit der päpstlichen Kanzlei im 12. Jahrhundert mit einem Blick auf den lateinischen Osten." In Herde, Peter, and Hermann Jakobs (eds.). *Papsturkunde und europäisches Urkundenwesen. Studien zu ihrer formalen und rechtlichen Kohärenz vom 11. bis 15. Jahrhundert.* Vol. 7. Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde. Köln: Böhlau, 1999. 1–26.

Jégou, Hervé, and Ondřej Chum. "Negative Evidences and Co-occurences in Image Retrieval: e Benefit of PCA and Whitening." In Fitzgibbon, Andrew et al. (eds.). *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, October 7-13 2012 Proceedings Part II.* Berlin: Springer, 2012. 774–787.

Jégou, Hervé, Florent Perronnin, Matthijs Douze, Jorge Sanchez et al. "Aggregating Local Image Descriptors into Compact Codes." *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 34.9 (2012). 1704–1716.

Li, Yuanpeng, Dmitriy Genzel, Yasuhisa Fujii, and Ashok C. Popat. "Publication Date Estimation for Printed Historical Documents Using Convolutional Neural Networks." *3rd International*

Workshop on Historical Document Imaging and Processing. Gammarth: ACM, 2015. 99–106.

LBA: *Lichtbildarchiv älterer Originalurkunden.* Philipps Universität Marburg. 2008. <http://lba.hist.uni-marburg.de/lba/>.

Lowe, David G. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60.2 (2004). 91–110.

Mom: *Collaborative Archive Monasterium.net.* 2005 –. <http://monasterium.net/mom>.

Nicolaou, Anguelos, Andrew D. Bagdanov, Marcus Liwicki, and Dimosthenis Karatzas. "Sparse Radial Sampling LBP for Writer Identification." *3th International Conference on Document Analysis and Recognition* (ICDAR). 2015. 716–720.

Papsturkunden.de. Göttingen Academy of Sciences and Humanities. 2015. <http://www.papsturkunden.de.

PuhMa: *Script and Signs. A computer-based analysis of high medieval papal charters. A key to Europe's cultural history.* 2012-2015. <http://www5.cs.fau.de/puhma>.

Su, Bolan, Shijian Lu, and Chew Lim Tan. "Binarization of Historical Document Images Using the Local Maximum and Minimum." *9th IAPR International Workshop on Document Analysis Systems*. 2010. 159–165.

Wahlberg, Fredrik, Lasse Mårtensson, and Anders Brun. "Large Scale Continuous Dating of Medieval Scribes Using a Combined Image and Language Model." *12th IAPR Workshop on Document Analysis Systems (DAS)*. 2016. 48–53.

Wahlberg, Fredrik, Lasse Mårtensson, and Anders Brun. "Large Scale Style Based Dating of Medieval Manuscripts." *Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing*. (HIP'15). 2015.

Some Roads to Script Classification: Via Taxonomy and Other Ways

Torsten Schaßan

Abstract

In codicology, the features of a script play an important role for dating and localising the manuscript. There are other questions that can be dealt with by examining these features, e.g. questions of intellectual history, influences of literary genres, or influences of organisational aspects of scriptoria on the shape of a script. But especially in the context of manuscript cataloguing the classification of script is of highest importance if other evidence such as a colophon or references like the naming of celebrations for local saints cannot be found. In order to contextualise the features of a script, palaeography has always striven for inference of a taxonomy from visual properties. Like in other disciplines, the community was not successful in achieving one common naming schema but constituted concurring taxonomies. Thus, the question arises what to do with these in times of the need to search huge amounts of manuscript related data in portals? New approaches in standardisation on the one hand, and semantic technologies and methods for image processing on the other hand, offer new possibilities to access to the manuscripts.

Zusammenfassung

In der Kodikologie spielt die Merkmale einer Schrift für die Datierung und Lokalisierung der Handschrift eine wichtige Rolle. Zwar lassen sich auch andere, geistesgeschichtliche Fragestellungen an diese Merkmale anknüpfen, wie etwa der Einfluss der Textsorte oder die Organisationsform eines Skriptoriums auf die Schriftgestalt, aber insbesondere im Kontext der Handschriftenkatalogisierung dient die Schrift dort, wo Kolophon, Nennung lokaler Heiliger oder andere inhaltliche Bezüge fehlen, der Ermittlung dieser wichtigen Information. Um die Merkmale einer Schrift in größere Bezüge einordnen zu können hat die Paläographie immer schon versucht, aus den visuellen Eigenschaften eine Klassifikation abzuleiten. Wie in anderen wissenschaftlichen Zweigen auch hat sich die Zunft aber nicht auf ein Benennungsschema einigen können, sondern konkurrierende Klassifikationen ausgebildet. Wie soll aber in einer Zeit, da verstärkt Handschriftenkataloge und andere handschriftenbezogene Informationen in Portalen durchsucht werden können und aufgrund der überwältigenden Menge auch durchsucht werden müssen, mit diesem Problem umgegangen werden?

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 165–177.

Neue Standardisierungsversuche auf der einen Seite, semantische Technologien und Bildverarbeitungsmethoden auf der anderen Seite bieten Möglichkeiten, Zugänge zu Handschriften zu ermöglichen.

1 Introduction

Today, a large amount of manuscript data is available from various digitisation efforts. How can these data be accessed? How do we find the way to a single manuscript or to a defined group of manuscripts? Besides other means such as the subjects covered, the works contained in, or persons related to, the manuscript, the classification of scripts may give researchers a tool to find the needle in the haystack. The main questions addressed in this paper are: How can such a classification be established? How has it been done in the past? What are the challenges and how can they be overcome with the possibilities of modern technologies and algorithms? While addressing these questions it shall be made clear that this article is not written from a palaeographer's point of view or claims to be a quest for some 'truth' about scripts and the names assigned, but from the perpective of someone who strives to support research by publishing manuscript data and who has to pave paths through masses of data, images, and descriptions alike. Thus, this paper focuses more on information theory and the usefulness or power of algorithms. Additionally, palaeography will be examined only in the context of script description and classification and not according to its possible other functions such as its relation to society and language or as an art (Cf. Stutzmann 2005, 16f.; Castro Correa 2014, 248).

2 What do we need script classification for, and why?

Stokes uses the term *palaeography* in the narrow sense as "the study of (medieval) handwriting with view towards its history and development and the identification, localization, and dating of *scribes*." (Stokes 2012, 137; emphasis by the author.) One would want to add: Palaeography is needed for localising and dating *manuscripts*. This task, the localisation and dating, has to start with 'basic truth', i.e. located and/or dated manuscripts, examine the script — and other, external evidences —, recognise its features and compare undated manuscripts with these examples in order to localise and date them. Now, this is what palaeographers have been doing since ages.

In order to share this information, e.g. through catalogues, the palaeographer had to describe what he/she saw in the manuscripts. Even today, catalogues describe scribal features. Derolez points out the problem with this approach: "How is it possible to proceed in such a way that the description of a specimen of handwriting is as clear

and convincing to its reader as it is to its author?" (Derolez 2003, 7) Firstly, author and reader have to use a shared language in order to understand each other, and, secondly, the reader has to know what to look for if he wants to recognise what the author has seen in a certain manuscript. And, as Derolez continues: "The method applied hitherto in palaeographical handbooks has produced an authoritative discipline, the pertinence of which depends on the authority of the author and the faith of the reader." (Derolez 2003, 9) This refers to the fact that the description establishes a special relation of belief, that the description in a catalogue is detailed enough to be understood and accurate enough to be true.

A very recent example for this problem is the following: Bernhard Bischoff is an authority if it comes to script, localising, and dating of manuscripts. When Hoffmann reviewed the last volume of Bischoffs *Katalog der karolingischen Handschriften*, edited by Birgit Ebersperger and published post-mortem, he criticised the editor for adding *Gudianus latinus 269* to the catalogue. She interpreted Bischoff and added Corvey as a place of origin. Hoffmann asks: "Und wer kann gar mit Sicherheit sagen, daß es Corveyer Hände waren? " (Hoffmann 2015, 17) It seems that Hoffmann would probably have trusted Bischoff but he scrutinises — and challenges — Ebersperger.[1]

Still, the long(er) descriptions of script have always been assigning a name to the script in question. This name represents the most common features of a certain script and is generalised from distinct hands. The term just offers a general impression of a script and does not allow for describing a certain hand. This term is listed in the indices of catalogues in order to allow for easy access to the manuscripts. The community learned to agree — more or less; we will come back to that — on a common list of terms.[2] The naming convention derived would be the basis for a controlled vocabulary in the first place and could be arranged into a taxonomy or even an ontology. Here, 'Digital Humanities methods' come into play, i.e. the application of technologies such as *RDF*, *TripleStores* and so on.

One remark on the notion of Stokes in his 2012 article on *Palaeography and the 'Virtual Library' of Manuscripts* in which he claims that "[t]he use of verbal description can probably never be avoided, because any use of a manuscript or facsimile is an act of interpretation." He continues: "We must be told which aspects of the letterforms are considered significant, how these significant differences compare between samples,

---

[1] Stutzmann claims that the status of being a connoisseur and, thus, an authority has long since been overcome and replaced by set objective criteria and precise terminology. A difference, however, would remain in the criteria applied. But, as Stutzmann critisised Derolez's system, by introducing 'accuracy' as another aspect to distinguish scripts, some level of subjective interpretation beyond the nomenclature continues to influence the analysis. (Stutzmann 2005, 19f)

[2] While Overgaauw had to conclude in 1994 that huge advances had been made for Carolingian and pre-Carolingian scripts but still no such advances were possible for the Gothic scripts (Overgaauw 1994, 100), Stutzmann reviewed the work of Derolez as "far more than just another palaeographic hand-book which offers a comprehensive nomenclature of gothic scripts." (Stutzmann 2005, 1)

and so on, and if we are not told this we are at sea as is demonstrated by an existing attempt to categorize letterforms by images alone." (Stokes 2012, 141) This quotation will have to be reviewed in the light of the techniques described in the last part of this paper as there are ways of measuring differences with digital technologies.

What one can find in all the data about scripts in the databases is the result of the reduction of specimen to simple, short terms. The collection of terms from indices forms the first approach to a controlled vocabulary. Yet, the community of cataloguers tends not to be satisfied with just a small number of terms and names for scripts but always strives for a better distinction between the scripts. A brief survey of recent cataloguing and digitisation projects highlights the problem to find a balance between the advantages of a very short list of terms versus a longer list of (probably) more accurate terms. Only such projects have been chosen which offer the cataloguer a predefined list of terms.[3] Table 1 lists the terms provided by the projects *ENRICH*, *Europeana Regia*, and the Swiss manuscript portal *e-codices*: The terms have been defined in the TEI schema for manuscript description. They are used at the elements `<handNote>` and `<scriptNote>`, more specifically on the attribute `@script` on these elements.

While there has only been defined a relatively small number of terms in the *ENRICH* project, the other projects add numerous terms to the list. All of the terms added are specific to the experiences made by the project partners respectively, representing the scripts that are common to the manuscripts in the collections or of that geographic region. During cataloguing those manuscripts, the participating institutions and heads of the projects must have felt the need to use these terms. The rather short list of the *ENRICH* project was meant to allow for searching and grouping the manuscripts by script in the first place. However, the other institutions and projects must have thought about a better representation of the heterogeneity of scripts. It is clear that too large a number of terms will serve none of the needs one might have: a list of too many entries will neither allow for faster access to the manuscripts, nor be able to describe the world of scripts and differences between hands and scripts in enough detail to replace imaging and the experience of the palaeographer. It is clear that, the closer one looks at scripts, the more differences one will recognise until not even two hands or scripts have the same properties in order to be called 'one script' or hand.[4]

---

[3] Not included are databases such as Manuscripta Mediaevalia which allow cataloguers to review the terms other cataloguers have used before them and just choose from them or enter any term they want. For that practice see Riecke 2009, 225: "Die Ansetzung der Eintragung sollte sich an den bereits vorhandenen Termini orientieren […]. Bislang wurden beispielsweise eingegeben: Buchschrift, gotische; Capitalis; Geheimschrift; Humanistica; Kanzleischrift; Kurrentschrift; Kursive; Majuskel, angelsächsische; Minuskel, karolingische; Perlschrift; Rotunda; Textualis formata; Unziale; Vortragsakzent."

[4] Cf. for that idea the 'Coastline paradox', which Peter Robinson applied to textual scholarship problems of textual variation in his paper (Robinson 1996. Original by Mandelbrot 1983).

ENRICH	Europeana Regia (added values)	e-codices (added values)
carolmin	capquad	antiqua
textualis	caprust	precar
cursiva	uncialis	spaetcar
hybrida	semiunc	praegot
humbook	benevent	gotica
humcursiva	luxeuil	semicursiva
kanzlei	corbie	greek[5]
kurrent	insulmin	hebrew
	alemmin	
	raetmin	
	carolgot	
	textura	
	rotunda	
	cancell	
	bastarda	
	cursant	
	cursrec	

Table 1: The terms provided by the projects *ENRICH*, *Europeana Regia* and *e-codices*

The general question would then be whether a community could agree upon a single list of terms for script classification at all? There are two kinds of problems connected to this: On the one hand, there is the difficulty to agree upon proper names for scripts that are similar to each other and might be distinguished only by minute characteristics. On the other hand, scripts might have been given different names over time although the visual features of those scripts would suggest likeness.

3 Traditional approaches

In order to examine the features of a script in larger contexts, palaeography has always striven to establish a taxonomy with inferences of the visual properties. Similar to other disciplines, the community was only partially successful to achieve one common naming schema, but constituted concurring taxonomies. These inherited different types of problems such as having different names for similar scripts, entities

[5] The inclusion of the terms 'greek' and 'hebrew' would add to the issues of classification of script discussed so far other aspects such as 'script and language'. As the topic of this paper is to examine the possibilities that certain technologies offer in order to overcome some problems, the inclusion of non-Latin scripts shall not be addressed any further.

overlapping geographically, or names of entities changing meaning over time causing ambiguity as one cannot be sure whether the term still covers the same entity, e.g. an area.

An example for the first issue, having different names for similar scripts, is the problem with terminology for Bastarda scripts as well as for Insular Carolingian minuscle. Stokes mentions this example of overlapping respectively divergent terminology. Alexander Rumble's guidelines include, among other terms, the 'round Anglo-Saxon minuscle'. This script is called 'Caroline minuscle' by Ker, 'Anglo-Insular minuscle' by Boyle, 'Anglo-Saxon round minuscle' by Brown, 'Anglo-Saxon vernacular minuscle' by Dumville, 'English Caroline minuscle' by Roberts, and, finally, 'English vernacular minuscle' by Stokes himself. (Stokes 2012, 147) One — traditional — way to deal with this issue would be to define a concordance and mention divergent names together, as e.g. Derolez does. (Derolez 2003; a concordance Stutzmann 2005, 63)

An example for the second problem, entities overlapping geographically and being not defined clearly, can be found in related subjects which are relevant for palaeography: the names of places and regions as used for localising script and manuscripts. When we find entries such as 'Südostdeutschland', 'Österreich', or 'Bayern' in a catalogue like Bischoff's catalogue of ninth centuries manuscripts, which entities do these refer to?

An example for the last issue, the change of names of entities over time, may be the distinction between 'Niederdeutschland' and 'Norddeutschland'.[6] Here, the same methodologies have to be applied.

4 'Healing' concurrence

At this point, it is necessary to repeat the definitions of some terms that are regularly — but sometimes perplexingly or wrongly — in the discussion of characteristics:

- A *controlled vocabulary* is just a collection of terms describing one aspect or feature. If the controlled vocabulary covers all aspects and is therefore 'complete', it is called a *nomenclature*.
- A *taxonomy* is an ordered, mono-hierarchical classification of the terms of a nomenclature.
- An *ontology* adds the relations between the terms to the mono-hierarchical classification.
- A *folksonomy* is a 'democratised' version of a taxonomy, derived from collaborative, social tagging.

[6] Examples taken from Hoffmann 2015, 45.

But how are any of these technologies applied? Cataloguers who write their documents in TEI-XML directly are offered lists of values for the classification of script during the cataloguing, as shown in table 1. The list is provided as part of a schema which defines the 'grammar' of a document. Cataloguers are provided with a template file that contains all the necessary structures and serves as a spreadsheet.[7] The template file references the schema file. The schema file contains the information about the scripts we want the encoders to specify and, respectively, the values of the attributes `@script` which can be used to name the script used in the manuscript. The list has been defined using the TEI *ODD* document type. *ODD* is short for 'one document does it all'. The *ODD* allows to define a schema from within the TEI. In this *ODD* file, the list of values is supplied. The definition of the list of values looks like this:

```
<classSpec ident="att.handFeatures" type="atts" mode="change" module="tei">
 <attList>
  <attDef ident="script" mode="change" usage="rec">
   <defaultVal>other</defaultVal>
   <valList type="semi" mode="replace">
    <valItem ident="carolmin">
     <desc xml:lang="de">Karolingische Minuskel</desc>
    </valItem>
    <valItem ident="textualis">
     <desc xml:lang="de">Textualis</desc>
    </valItem>
    <valItem ident="cursiva">
     <desc xml:lang="de">Kursive</desc>
    </valItem>
    <valItem ident="hybrida">
     <desc xml:lang="de">Hybrida</desc>
    </valItem>

    <!-- more values might be defined here -->
    <valItem ident="other">
     <desc>any other type of script</desc>
    </valItem>
    <valItem ident="unknown">
     <desc>script information not available</desc>
    </valItem>
   </valList>
  </attDef>

    <!-- more attributes might be dealt with here -->
 </attList>
</classSpec>
```

A list like this represents a controlled vocabulary. The cataloguers will be able to choose from this list of terms during their cataloguing. Managing the list helps to foster the interoperability of data and allows for faceted browsing of the catalogue entries if implemented. Lists like these are helpful especially if the list of terms included is rather short.

---

[7] <https://github.com/schassan/cataloguing/blob/master/tei-msDesc_template.xml>.

If one wanted to allow for a better search even with a large number of entries, one could arrange the terms in a mono-hierarchical classification, a taxonomy. In the resulting hierarchy, one would distinguish broader and narrower terms, e.g. *Cursiva* as broader and *Cursiva antiquior* or *Cursiva recentior* as more specialised terms. Furthermore, concurring terms like the ones mentioned above could remain as they are whilst their relation with others can be expressed. To express a hierarchy of or relations between terms, one could apply several semantic web technolgies: one of these is the usage of the *Web Ontology Language* (OWL).[8] The technical realisation the hierarchy for the example in *OWL* might look like that:

```
<rdf:RDF>
 <owl:Ontology rdf:about=""/>
 <owl:Class rdf:ID="script"/>
 <owl:Class rdf:ID="cursiva" rdf:about="http://anyuri.com/scripts#cursiva">
  <rdfs:subClassOf rdf:resource="#script"/>
  <rdfs:label>Cursiva</rdfs:label>
  <rdfs:comment>This class covers all types of cursive script.</rdfs:comment>
 </owl:Class>
 <owl:Class rdf:ID="cursant">
  <rdfs:subClassOf rdf:resource="#cursiva"/>
  <rdfs:label>Cursiva antiquior</rdfs:label>
  <rdfs:comment>This class covers antique cursive script.</rdfs:comment>
 </owl:Class>
 <owl:Class rdf:ID="cursrec">
  <rdfs:subClassOf rdf:resource="#cursiva"/>
  <rdfs:label>Cursiva recentior</rdfs:label>
  <rdfs:comment>This class covers recent cursive script.</rdfs:comment>
 </owl:Class>
</rdf:RDF>
```

This example implements the ontology in *RDF* syntax and therefore uses elements from the *RDF Schema* (RDFS) namespace. RDFS offers elements to define sub- and super-classes as well as relations between classes such as ‘sameAs’, ‘similarTo’, or ‘relatedTo’. With the means of semantic web technologies and *RDF* it is possible to enhance a controlled vocabulary respectively a taxonomy to an ontology.

5 Machine-aided approaches

Already in 1979, Bernhard Bischoff recognised that palaeography, which used to be an art of vision and empathy, becomes an art of measuring by technical means.[9]

---

[8] Although the abbreviation for the Web Ontology Language should be *WOL*, the reason for choosing *OWL* is not entirely clear. Possible explanations include that the inventor of that language chose to introduce a more interesting one, stating that “Why not be inconsistent in at least one aspect of the language which is all about consistency?” (Schreiber) Another one is that this acronym has been chosen as a tribute to William A. Martin’s *One World Language* knowledge representation project from the 1970s.

[9] “Mit technischen Mitteln ist die Paläographie, die eine Kunst des Sehens und der Einfühlung ist, auf dem Wege, eine Kunst des Messens zu werden.” (Bischoff 1986, 19)

In a 'machine-aided approach', the traditional palaeographic method is enhanced by the aid of some automated methods, e.g. measuring.[10] Already in 1977, Gilissen pioneered with the statistical analysis of quantitative measurements such as pen-angle, pen-width, etc. (Gilissen 1977, cited in Stokes 2012, 145)

An example for a machine-aided approach to palaeographical research is the project DigiPal. DigiPal has been developed at the University College London by Peter Stokes et al. The basis of DigiPal is a database to which researchers can add images of manuscripts and detailed descriptions such as own characterisations of scripts and classifications and other metadata. Users can cut out single letters which then will be displayed side-by-side with other occurrences of the same letter. This generates collections which can be grouped, compared, and searched for. All this work has been and is to be done manually. Moreover, as there are only single letter-forms stored in this database, they lack the context of the word, line, and entire page. This is exactly what the algorithms presented afterwards are going to look at. The machine-aided part of the project consists of a set of search algorithms which are able to search both in the descriptions of script as well as in the descriptions of characteristics. The characteristics of script can be detailed enough to serve as a finger-print of a script. These fingerprints are compared by the algorithms.

Aussems and Brink presented another possibility by looking at a "writer-specific variation in the width of the ink trace" and measuring "the relation between the local direction and width of the ink traces." (Aussems and Brink, 298)

As Stokes points out, "[…] none of these projects accounts for page curvature, image distortion, or the natural expansion and contraction of parchment […] " (Stokes 2012, 145). I think he is exagerating here as the human eye is subject to the same challenges and the palaeographer's decisions have to be questioned as well. The algorithms mentioned above do not account for that either and may have to be adjusted in order to do so. On the other hand, as measurements are summarised over many pages, sometimes entire manuscripts, the deviation may be of little significance.

6 Machine-driven approaches

To overcome the burden to describe every detail in one's own words, only some printed catalogues supply the reader with a series of images from the manuscripts. This is especially true for catalogues of dated manuscripts which contain both images of pages that have a colophon or other means used for dating, and images of sample pages representative of the script used throughout the manuscript. Catalogues of illuminated manuscripts supply images for art historical means. But even 'normal'

[10] Stokes called this approach 'computer-aided' and gave an overview of the questions and methods in his 2009 publication.

catalogues sometimes supply a number of images.[11] With these images, the reader can make up his mind and compare the description of script with its actual image. Still, *an image says more than a thousands words.*

But even today, as more and more images of manuscripts are available online for consultation and reference, one would want to access this huge amount of data with the help of standardised terminology or via a pattern that can be found in all these images. This time, we do not need the terminology for summarising the long feature descriptions but in order to subdue the sheer mass of information available to us. To supply a term or a reference to a pattern for every image available will have to be the task of (automated) algorithms, in the best of all worlds implemented as services.

Such algorithms have been proposed e.g. by Bulacu and Schomaker (2007a and 2007b), others built on top of these basic algorithms. (Cf. Fecker et al. 2015) Basically, these algorithms are based on the idea that a script can be described as a multidimensional matrix of attributes such as stroke-width, slant, etc. Once all of these characteristics have been recognised, measured, and assessed, the algorithm is (or should be) able to distinguish between different scripts. Although the algorithms mentioned above have been used for scribal identification and are, thus, intended to find differences in what are supposed to be similar or homogenous measurements, one would think that the difference between scripts — in order to arrive at a classification — e.g. between Caroline minuscule and Gothic scripts, should be greater than the differences between two hands writing both a Gothic minuscule? Another aspect would be that the proposed algorithms strive both for a script identification as well as for a script verification. The identification will separate the characteristics of a script from possibly all other scripts. This does not necessarily imply a writer identification which would mean to assign an identified script to an identified scribe. But once the script has been identified by recognising its features, it should be possible to look for the same features in other manuscripts or on other pages and, thus, verify whether a page or a manuscript has been written by the same 'hand'.

It has to be stressed that importance of certain attributes of script varies if one examins scribal hands or scripts. Finding and defining attributes that scripts have in common and then have an algorithm to process the image data might be as complicated as it is for palaeographers to agree on a common terminology.[12]

---

[11] Whether (text-)catalogues contain images or not seems to depend more on money than on a theory behind their establishment. Thus, catalogues without images are more frequent but there are a number of catalogues that contain sample images, cf. the catalogues of Jena."

[12] For further discussion of the issues cf. Stutzmann 2015.

7 Conclusions, or: Are algorithms the better palaeographers?

The answer to that question depends on what you want the palaeographer to do. The machine is definitely capable of recognising features of scripts. The algorithms can calculate means of pretty much everything: thickness, straightness, or orientation of strokes, height or width of lines, numbers of lines per page, etc. From those general features of script, the algorithm can determine clusters. Depending on thresholds, the machine is able to distinguish individual scripts and maybe even script families. Whether it is possible to attribute terms to these clusters, or if the algorithm ends at the same position as palaeographers did, is an open question.

Anyway, the ultimate question might be: to what end do we use the classification? One option might be to find as many objects as possible that meet one's criteria in the course of a research project in order to analyse them and answer research questions. The other might be to find the 'correct' items in a given set of objects. This would require much higher 'accuracy'. Institutions such as libraries may be more interested in the first option: one needs to find out about similar objects which have to be examined by experts in order to help them with their cataloguing or their research.

Is there anything like being 'right' or 'wrong' when talking about script or, more general, about palaeography?

Most importantly, the history of script(s) is no mono-hierarchical development, emerging from Capitalis via Uncialis to Minuscules, to name but a few. Thus, the technical means to deal with the phaenomenon of scripts could be the implementation and application of a taxonomy, but this would represent a mono-hierarchy. The better way would be to establish an ontology. In order to derive the net of taxonomies, the algorithms presented can be used to generate a basic knowledge.

One has to conclude an overview like the one above with the *almighty Bernhard Bischoff*. Hoffmann cites him as follows: "Berühmt, um nicht zu sagen berüchtigt, ist seine [Bischoffs] Charakterisierung des Reimser Stils: das lange s sei dort stärker geneigt gleich Getreidehalmen, über die der Wind gehe." *Famous, not to say notorius, is his [Bischoff's] characterisation of the Reims style: the long s were more slanted there like the stem of grain in the wind.* (Hoffmann 2015, 40) Not to be left with such wonderfully poetic yet hard to comprehend descriptions may be the task of the new methods.

Bibliography

Aussems, Mark and Axel Brink. "Digital Palaeography." In Rehbein, Malte, Patrick Sahle, and Torsten Schaßan (eds.). *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age.* Norderstedt: Books on Demand, 2009. 293–308. URN: urn:nbn:de:hbz:38-29773.

Bischoff, Bernhard. *Paläographie des römischen Altertums und des abendländischen Mittelalters*. Second revised edition. Berlin: Schmidt, 1986.

Bulacu, Marius L. and Lambert R. B. Schomaker. [2007a.] "Text-independent Writer Identification and Verification Using Textural and Allographic Features." *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Special Issue - Biometrics: Progress and Directions* 29.4 (2007). 701–717.

Bulacu, Marius L. and Lambert R. B.Schomaker. [2007b.] "Automatic handwriting identification on medieval documents." *Proceedings of 14th International Conference on Image Analysis and Processing (ICIAP, Modena, 11–13 September 2007)*. Los Alamitos: IEEE, 2007. 279–284.

Castro Correa, Ainoa. "Palaeography, Computer-Aided Palaeography and Digital Palaeography: Digital Tools applied to the Study of Visigothic Script." In Andrews, Tara and Caroline Macé (eds.). *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*. Turnhout: Brepols, 2014. 247–272.

Derolez, Albert. *The palaeography of Gothic Manuscript Books from the Twelfth to the Early Sixteenth Century*. Cambridge: Cambridge University Press, 2003.

DigiPal: *Digital Resource and Database of Palaeography, Manuscript Studies and Diplomatic*. 2011-2014. <http://digipal.eu>.

e-codices. Virtual Manuscript Library of Switzerland. 2005-. <http://www.e-codices.unifr.ch>.

ENRICH: *ENRICH project*. University of Oxford. 2007-2009. <http://enrich.manuscriptorium.com>.

Europeana Regia. 2010–2012. <http://www.europeanaregia.eu/>.

Fecker, Daniel, Volker Märgner, and Torsten Schaßan. "Vom Zeichen zur Schrift: Mit Mustererkennung zur automatisierten Schreiberhanderkennung in mittelalterlichen und frühneuzeitlichen Handschriften." Baum, Constanze, and Thomas Stäcker (eds.). *Grenzen und Möglichkeiten der Digital Humanities*. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). 2015. DOI: 10.17175/sb001_008.

Gilissen, Léon. *Prolgomènes à la codicologie: Recherches sur la construction des cahiers et la mise en page de manuscrits médiévaux*. Ghent: Éditions scientifiques, 1977.

Die Handschriften der Thüringer Universitäts- und Landesbibliothek Jena. Vol. 1–3. Wiesbaden: Harrasowitz, 2002–2016.

Hoffmann, Hartmut. "Bernhard Bischoffs Katalog der karolingischen Handschriften." *Deutsches Archiv* 71 (2015). 1–56.

Mandelbrot, Benoît. "How Long Is the Coast of Britain." *The Fractal Geometry of Nature*. New York (NY): W.H. Freeman and Co., 1983. 25–33.

ODD: *One document does it all*. Meta-schema file defined by the TEI which allows to generate schema files in multiple schema languages. <http://www.tei-c.org/Guidelines/Customization/odds.xml>.

Overgaauw, Eef. "Die Nomenklatur der gotischen Schriftarten bei der Katalogisierung von spätmittelalterlichen Handschriften." *Codices manuscripti* 15 (1994) 100–106.

RDF: *Resource Description Framework*. World Wide Web Consortium (W3C). 2014. <http://www.w3.org/RDF>.

RDFS: *RDF Schema 1.1*. World Wide Web Consortium (W3C). 2014. <https://www.w3.org/TR/rdf-schema>.

Riecke, Anne-Beate. *MXML-Dokumentation. Die Erstellung von Handschriftenbeschreibungen nach den Richtlinien der DFG mit Hilfe von Manuscriptum XML*. Berlin: Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, 2009.

Robinson, Peter. "Is there a Text in These Variants?" In Richard Finneran (ed.). *The Literary Text in the Digital Age.* Ann Arbor (MI): University of Michigan Press, 1996. 99–115.

Schreiber, Guus, cited by Ivan Herman. *Why OWL and not WOL?*. Tutorial on Semantic Web Technologies. World Wide Web Consortium. <https://www.w3.org/People/Ivan/CorePresentations/RDFTutorial/Slides.html#(114)>.

Stokes, Peter. "Computer-Aided Palaeography, Present and Future." In Rehbein, Malte, Patrick Sahle, and Torsten Schaßan (eds.). *Kodikologie und Paläographie im digitalen Zeitalter - Codicology and Palaeography in the Digital Age.* Norderstedt: Books on Demand, 2009. 309–338.

Stokes, Peter. "Palaeography and the 'Virtual Library' of Manuscripts." In Nelson, Brent, and Melissa Terras (eds.). *Digitizing medieval and early modern material culture.* Toronto: ACMRS, 2012. 137–169.

Stutzmann, Dominique. "Nomenklatur der gotischen Buchschriften: Nennen? Systematisieren? Wie und wozu? Recension of: Albert Derolez: The Palaeography of Gothic Manuscript Books. From the Twelfth to the Early Sixteenth Century. Cambridge et. al.: Cambridge University Press 2003.)." *IASLonline.* 2005. <http://www.iaslonline.de/index.php?vorgang_id=995>.

Stutzmann, Dominique. "Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol". *Digital Medievalist* 10 (2015). DOI: 10.16995/dm.61/.

TEI: *Text Encoding Initiative.* <http://www.tei-c.org>.

Phenetic Approach to Script Evolution

Gábor Hosszú

Abstract

Computational palaeography, as a branch of applied computer science, investigates the evolution of graphemes, explores relationships between scripts, and provides support for deciphering ancient inscriptions, among others. The author applied methods often used to describe evolutionary processes in phylogenetics to analyse the development of scripts. Unlike in the clear evolution of phylogenetics, graphemes used to describe the evolution of scripts are sometimes indistinguishable from their glyph variants. Moreover, the historical background is at times incomplete. In order to reduce uncertainty, the author developed an exploratory data analysis method that combines phenetic analysis methods with a cladistic approach. The paper details the tests the author developed to explore the relationships among 66 different scripts with 186 different features. To extract data for analysis required determining the similarity groups of glyphs and orthographical rules in different scripts; the input is data from humanities-based palaeography. Creation of the similarity groups of the glyphs is based on minimizing the differences between the topological properties of the glyphs and individual decisions in order to avoid homoplasies, as well as the erroneous omission of slightly differing but otherwise related glyphs. For the second purpose, the layered grapheme model and the concept of characteristic transformations of related glyphs were used. Based on the extracted features of the scripts, various machine-learning methods were applied, including multidimensional scaling, k-means partitional clustering, and various hierarchical clustering methods. These algorithms produced similar results, represented in two- and three-dimensional scatter plots and phenograms, which visualize the relationship between the scripts. These results roughly concur with the results of humanities-based palaeography; however, new conclusions can be also derived, including the introduction of the concept of witness scripts, and glyph- and grapheme-level reticulations, which are used to describe the possible relationship of graphemes and scripts. The presented results demonstrate the usefulness of a developed modified phenetic method in exploring the similarities of scripts, and based on the results obtained, some improvements in modelling the distribution of certain historical scripts were also proposed.

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 179–252.

Zusammenfassung

Computergestützte Paläographie als Zweig der angewandten Informatik untersucht unter anderem die Evolution von Graphemen, erforscht die Beziehungen zwischen Schriften und leistet Unterstützung bei der Entzifferung sehr alter Inschriften. Der Autor hat Methoden, die häufig für die Beschreibung evolutionärer Prozesse verwendet werden, angewandt, um die Entwicklung von Schriftsystemen zu untersuchen. Im Gegensatz zu der klaren Evolution in der Phylogenetik, sind Grapheme, die zur Beschreibung der Schriftevolution benutzt werden, manchmal nicht von ihren Glyph-Varianten zu unterscheiden. Zudem ist der historische Hintergrund zuweilen unvollständig. Um die Unsicherheiten zu reduzieren, hat der Autor eine explorative Methode der Datenanalyse entwickelt, die phänetische (numerisch taxonomische) Analysemethoden und einen kladistischen Ansatz kombiniert. Der Beitrag erläutert die Testreihen, die der Autor entwickelt hat, um die Beziehungen zwischen 66 verschiedenen Schriften mit 186 verschiedenen Merkmalen zu erforschen. Die Datenextraktion für die Analyse machte es notwendig, zunächst die Ähnlichkeitsgruppen von Glyphen und die orthographischen Regeln für verschiedene Schriften zu bestimmen; die Ausgangsdaten stammen also aus der traditionellen Paläographie. Die Bestimmung der Ähnlichkeitsgruppen basiert sowohl auf der Minimierung der Unterschiede zwischen den topologischen Eigenschaften der Glyphen und individuellen Entscheidungen zur Vermeidung von Homoplasien (zufälligen Ähnlichkeiten), als auch der falschen Aussonderung von nur leicht unterschiedlichen, ansonsten aber ähnlichen Glyphen. Für die zweite Aufgabe wurden das Graphem-Schichtenmodell und das Konzept der charakteristischen Transformationen verwandter Glyphen benutzt. Auf der Grundlage der bestimmten Merkmale wurden verschiedene Methoden des maschinellen Lernens wie multidimensionale Skalierung, k-Means Partitions-Clusteranalyse und verschiedene hierarchische Clusterverfahren angewandt. Diese Algorithmen haben zu ähnlichen Ergebnissen geführt, die in zwei- und dreidimensionalen Streudiagrammen und Phänogrammen (Kladogrammen) ausgedrückt werden und die Verhältnisse zwischen Schriften sichtbar machen. Die Ergebnisse stimmen grob mit den Resultaten der bisherigen paläographischen Forschung überein, allerdings können aus ihnen auch neue Erkenntnisse gezogen werden. Dazu gehören die Einführung des Konzepts der »Zeugenschriften« und Verbindungen auf der Glyph- und Graphemebene, die zur Beschreibung möglicher Beziehungen zwischen Graphemen und Schriften genutzt werden. Die hier vorgestellten Ergebnisse zeigen den Nutzen einer entwickelten phänetischen Methode für die Untersuchung von Schriftähnlichkeiten. Auf der Grundlage der erzielten Resultate werden außerdem Verbesserungsvorschläge für die Modellierung der Verbreitung und Verteilung einiger historischer Schriften gemacht.

1 Introduction

Computational palaeography, in other words *engineering in palaeography*, as a branch of applied computer science, deals with investigating the evolution of graphemes, exploring relationships between scripts, and providing support for deciphering ancient inscriptions, among others. Its main focus is using engineering methods to explore relationships found in the data of ancient inscriptions and other palaeographical (including epigraphic) information. Computational palaeography has an applied machine learning approach, and it extends the engineering modelling methods to any data of the written cultural heritage. The fields of computational palaeography are improving and tailoring phylogenetic algorithms for exploring relationships in palaeographical data and modelling the evolution of scripts and graphemes, including the spatial analysis of the various glyphs. The research efforts of the author and his colleagues cover a broad range of topics such as applying machine learning methods to explore similarities among scripts or orthographies (Hosszú 2014; Tóth et al. 2016), modelling graphemes in different abstraction levels (Pardede et al. 2016), reconstructing lineages of graphemes in various scripts (Hosszú 2015), investigating methods for testing the appropriateness of the reconstructed lineages, and developing algorithms for deciphering historical inscriptions (Tóth et al. 2015).

As opposed to computational palaeography, *digital palaeography* (Ciula 2005; 2009)—or in other words *computer-aided palaeography* (Stokes 2009)—is part of Digital Humanities, an interdisciplinary field of Palaeography, Computing, and Artificial Intelligence (Aussems and Brink 2010, 296). It is an extension of the type of palaeography found in the humanities using tools from computer science; their goals are similar (e.g. Aussems 2010). Humanities-based palaeography, with diplomatics and textual criticism, constitutes the main disciplines of philology. For simplicity, the term palaeography includes epigraphy in this article. Digital palaeography includes sub-fields such as *quantitative codicology* (Stokes 2015) and *quantitative palaeography*, and it entails the identification of scribes, reconstruction of fragmented texts with image analysis, digital representation of medieval scripts, digital description, imaging, recording, and reproduction of the manuscripts, image pre-processing for machine learning (e.g. feature extraction, pattern recognition, optical character recognition), textual analysis, physical analysis, storage in databases extending with semantic structures, digital presentation, and the teaching of palaeography (Ciula 2009; Fischer et al. 2010). Quantitative aspects can be measured by automated means and the results can be subjected to automated clustering techniques (Ciula 2005). Hierarchical clustering was used for creating the groups of the morphologically similar glyphs of a grapheme. A composite palaeographical classification method, including k-means clustering, was applied to match a particular document to a large set of palaeographical records (Wolf et al. 2011). Numerical tools were developed to automate the study of medieval

writing samples in the context of the Graphem project, which is intended to explore, analyse, and categorize medieval scripts (Cloppet et al. 2011). It is noteworthy that the border between digital palaeography and computational palaeography is smooth, both of them use machine-learning tools, and they are related to analytical palaeography, which deals with the classification of glyphs and belongs to the palaeography of the humanities.

This paper details the investigations carried out to explore the relationships among 66 different scripts using clustering and factor analysis, where 186 different features of the examined scripts were involved in the phenetic analysis. As the input of the analysis, a data extraction step is necessary, which means determining the similarity features groups (SFGs) in different scripts; where the input was the result of humanities-type palaeography, and the criteria for constructing the SFGs include phenetic and cladistic considerations. Various machine-learning methods were applied, including multidimensional scaling, k-means partitional clustering, and different hierarchical clustering methods, and the different algorithms produced similar results; they are represented in two- and three-dimensional scatter plots and phenograms, where each point represents a single script, and the relative distance of these points represents the relationship between the scripts.

The paper is organized as follows: Section 1 gives background information, including a definition of the concepts and terminology of machine learning, comparison of phylogenetic approaches, details of phenetic tools, cluster validity techniques, and the terms and concepts of computational palaeography. Section 2 is dedicated to the newly developed exploratory data analysis method, including the general description of the algorithm. Section 3 presents the feature extraction with SFGs, section 4 evaluates the obtained results, and section 5 provides conclusions. A short Appendix presents some additional examples of the inscriptions written with the lesser-known scripts of the Eurasian Steppe.

2 Background

2.1 Computational palaeographical and machine learning terminology

A *writing system* is "a set of visible or tactile signs used to represent units of language in a systematic way" (Coulmas 1999). *Script* is the graphic form with orthographical rules of a writing system. A script has several versions, including the subset of the graphemes of the script belonging to various areal, cultural, temporal, stylistic, and typographical versions. An extinct script, for which only inscriptions have survived, is reconstructed from the surviving inscriptions, including their explored properties (e.g., orthographical rules).

Orthography is a certain set of the graphemes of a script and a set of rules for using a script in a particular language; e.g., some medieval orthographies of the Latin script include medieval German, medieval Italian, Middle English, Old French, Old Hungarian, Old Norse, etc. In computational palaeography, the term orthography means a specific set of graphemes with specific glyphs; e.g., the ê in French and medieval Italian, the ß in German, the ȝ in Old English, the ÿ in Dutch, French, medieval German, and Old Hungarian, ǫ in Old Norse and Old Hungarian orthographies, etc. All of these graphemes belong to the Latin script.

Taxon is a taxonomic unit, a set of objects classified into the same category in a formal taxonomic system. In biological evolution, taxa are usually species, and the entities of the species are called organisms. In computational palaeography, the taxa are the scripts, and the entities are the particular versions of a script (orthographies) used for each inscription. However, other approaches are also possible depending on the focus of the research – if the broad focus is on a particular orthography, it could be considered a taxon and variations of its graphemes would be the entities.

Grapheme is the smallest semantically distinguishing element in a script (Sukkarieh et al. 2012). A grapheme could be a letter, ideogram, logogram, ligature, numerical digit, diacritic, accent, phonogram, determinative, punctuation mark, syllabogram, etc. The grapheme is taken as an object with different features including its shape variations (called glyphs), transliteration values, sound values, age in which it was used, geographical distribution area, and the script to which it belongs.

Glyph refers to a unique shape of a grapheme that can be described by topological information. In the view of computational palaeography, the definition of a grapheme has the following conjunctive constituents: *(i)* different phonemes belong to the same grapheme if the sets of their possible phonetic values are identical or reasonably altering; *(ii)* the glyph variants of a certain grapheme must be visually very similar; *(iii)* any glyph variants of a certain grapheme must represent all phonetic values of that grapheme; and *(iv)* the usage of the grapheme is determined by the orthographical rules of a certain age in the history.

Inscription is a survived relic of one or more scripts independent of the writing materials (stone, wall, wood, ink and paper/papyrus/parchment, etc.), and physically it can be a fragment, a manuscript, a scroll, or a codex. In other words, the term inscription is used in the widest possible sense.

Symbols are the minimum individual units of the inscriptions from a visual perspective. In other words, inscriptions are composed of a sequence of symbols. Consequently, a symbol is the materialization of a particular glyph of a grapheme, and the grapheme is the abstraction of a symbol.

Feature (also called *character*) is a heritable trait (property) of a taxon. A feature can take one of more forms; these various forms are described by the feature states. It is

noteworthy that in phylogenetics, the term "character" is much more frequently used than the term "feature"; however, in pattern recognition, the term "character" is used very similarly to the term "grapheme." In computational palaeography, the concepts of both phylogenetics and pattern recognition are used, which makes the term "character" ambiguous. Therefore, instead of the ambiguous term "character," the preferred terms are "feature" and "grapheme," respectively. A computational palaeographical feature is any property of scripts that can take one or more forms; these different forms are called states of the features. These features could be graphemic and orthographic. *Graphemic features* are represented using a binary variable having the two states "presence of a glyph variant of a grapheme" and "absence of a glyph variant of a grapheme". Similarly, *orthographic features* represent the presence or absence of various orthographic rules, e.g., directions or separator lines among rows of the inscriptions of certain scripts. In other words, the categorical variables are transformed into Boolean indicator variables (see below for details).

Object is the basic unit (data point) in machine learning methods, which is described with a vector of *variables* (in other words, *attributes*). If the machine learning method is used in phylogenetics, the object is the taxon, and the variable is the feature. Therefore, the object is usually the script, and the variable is the feature of the script, especially the existence of certain glyphs in the given script. In such a way, the *taxon-feature matrix* is composed of taxons in rows, and feature states in columns. If the feature is transformed into a Boolean indicator variable, the value of a feature state is 1 if it is present in a particular taxon, and 0 if it is absent.

Clade is a taxon and all of its descendant taxa (Hennig 1966). The taxa in a single clade share an evolutionary relationship. The taxa have features, and a taxon can be characterized by the feature states. *Apomorphy* is a derived feature state of a taxon; this feature state is known as *apomorphic*, and includes the types called autapomorphy, synapomorphy (homology), or homoplasy (analogy). *Autapomorphy* means a feature is present in an individual taxon, but not any of its ancestors. If there are descendants of a taxon that inherit this autapomorphic feature, then they create a clade, and this clade is characterized by this feature as apomorphy. This demonstrates that the properties apomorphy or autapomorphy are relative terms. *Synapomorphy* is a feature state shared by two or more taxa resulting from an innovation in their shortest common ancestor. Synapomorphy is a *homology,* meaning a similarity due to inheritance of a feature state from a common ancestor. *Homoplasy* (homoplastic feature state) is when two or more apomorphic feature states are identical; however, they originated from not a common ancestor, but rather by convergence or reversal. *Convergence* (parallel evolution) is when the same feature state presents in two unrelated taxa due to similar conditions. *Reversal* (back-mutation) is when a feature state reverts to an earlier state. *Plesiomorphy* is a feature state that taxa of a clade

have retained from their ancestors; such feature state is *plesiomorphic* (ancestral). Considering a clade, the common feature states of its taxa may be plesiomorphic or synapomorphic.

Reticulate evolution happens in the case of *reticulate* events (Sneath 1975), such as hybridization (a new taxon is formed from two different taxons) or horizontal gene transfer (feature state transfer). In computational palaeography, *hybridization* means the combination of two scripts, e.g., the Early Cyrillic script is surely a hybrid, or combination, of the Greek and the Glagolitic scripts. *Feature state transfer* means transfer of a glyph or orthographical rule between contemporaneous scripts. It has two subcases: *(i) Glyph-level reticulation:* if a grapheme exists in a script, but an additional glyph for this grapheme is transferred (borrowed) from another script (loan glyph); or the grapheme did not exist in a script, and its glyphs are transferred from more than one grapheme (new grapheme with loan glyphs). See the comment on SFG-100 in table 10 for an example. *(ii) Grapheme-level reticulation:* if a grapheme with all of its glyphs in a script originated from a certain grapheme of another script (loan grapheme). For instance, the Latin script was developed from the Greek-origin Etruscan graphemes, and, later the graphemes Y and Z were directly adopted from the Greek. In this example, the graphemes were absent from the Latin script at the time of adoption.

Witness script is a script that has retained features from another script from a remote geographical region and/or a bygone era. For example, the Greek script retained features from an early form of the Phoenician script. Thus we can say that the Greek script bears witness to certain features existing in the early Phoenician script.

2.2 Phylogenetic approaches for computational palaeography

Phylogenetics aims to uncover the evolutionary relationships between taxa to obtain an understanding of their evolution. Phylogenetics in a wider sense has three areas: *phenetics* (numerical taxonomy), *cladistics,* and *phylogenetics* (in a narrow sense). The output of these methods is usually presented in tree-like branching diagrams (dendrogram, indexed tree) called *phylogenetic trees* (in a wider sense), or the *phenogram, cladogram,* and *phylogram* (in a narrow sense), respectively. A tree is a connected acyclic graph consisting of a set of vertices (nodes) and a set of edges (branches), each of which connects a pair of vertices. The differences between phenograms, cladograms, and phylograms are related to their underlying features: phenograms use phenotypic information, cladograms use hierarchical relationships among taxa based upon homologies (synapomorphies), and phylograms convey genealogical information. The phenetic relationships are usually multidimensional; therefore, different procedures can produce a variety of phenograms (Sokal et al. 1963). The cladogram is

a synchronic representation of the evolution; it describes the relationships among the taxa of the same time. A phylogram is an estimation of genealogical relationships among a group of taxa (Kitching et al. 1998, 213); it represents evolutionary histories in which the main events are speciations (at the internal nodes of the tree) and descent with modification (along the edges of the tree).

The lengths of the branches of the tree have different meanings in the three approaches. In phenograms the length of the branch represents the similarity among the taxa. In cladograms, the length of the branch has no specific meaning. In phylograms, the length of the branch represents the amount of inferred evolutionary change: the longer the branch, the greater the variation between taxa. If in a tree, two scripts have a more recent common ancestor, then we expect these two scripts to have the most features in common, because they are the pair that has had the least opportunity to diverge. Using more than one feature provides a measure of the overall difference between them. It is assumed that the features in common are not convergent and have not evolved independently in the two branches by chance, either. The differences likely accumulate at a fairly steady rate, so that more differences mean that there is a less recent common ancestor.

The tree-based phylogenetic model is less suited for reticulate events, when the new taxon has more than one ancestor. In this case, *phylogenetic networks* better describe the evolution than phylogenetic trees. When scripts converge in the case of reticulate evolution, a network model is more appropriate with additional edges to reflect the dual parentage of a script. These edges could be bidirectional if both scripts borrow from one another. Change happens continually to scripts, but not usually at a constant rate, with its cumulative effect producing splits into orthographical variations and script families. Finally, there could be loss of any evidence of relatedness. Unlike biology, it cannot be assumed that scripts all have a common origin; relatedness must be established.

An analogue of the field of computational palaeography is the use of phylogeny for historical linguistics (Forster and Renfrew 2006). Methods of computational phylogenetics and cladistics can be used to define an optimal tree or network to represent a hypothesis about the linguistic evolution. Nakhleh et al. (2005) compared the following phylogenetic reconstruction methods on an Indo-European linguistic dataset: UPGMA, maximum parsimony, weighted and unweighted maximum compatibility, neighbour joining, and Gray-Atkinson algorithms. They found that UPGMA is inferior on these datasets, because they used data from different time depth. The other algorithms were not sensitive to this feature of the applied datasets. The maximum parsimony and the unweighted maximum compatibility methods returned similar dendrograms. Their dataset for the phylogenetic reconstruction for comparative historical linguistics contains lexical, phonological, and morphological features. Bar-

Issue	Species	Languages	Scripts
Reticulation	Possible (e.g., horizontal gene transfer)	Possible (e.g., loan words)	Possible (hybridization of scripts, glyph-level reticulation, grapheme-level reticulation)
Interrelation of features	Frequent (biological feature states usually affect each other)	Possible (shared cultural development)	Possible (e.g., influence of the writing techniques, or effect of geometric style)
Homoplasy	Possible (e.g., parallel evolution)	Rare, historical linguists can identify many of the borrowings; therefore, they can be screened out (Barbançon et al. 2013)	Frequent (e.g., presence of similar, but unrelated glyph variants in unrelated scripts)

Table 1: Comparison of biological, linguistic, and palaeographical feature evolutions

bançon et al. (2013) used Hamming distances in the UPGMA to define the distance matrix between the set of languages.

Another interesting application of the phylogenetic approach is the examination of the evolutionary relationships in software (Sampaio 2007). However, in the cases of modelling software evolution and biology, different methods are used to compare and classify the taxa.

Phylogenetic reconstruction methods originally designed for biological data could be used on palaeographical data for reconstruction of the phylogenies of script families (table 1). Analysis can be carried out on the features of scripts. Each feature being a Boolean indicator defines an equivalence relation on the set of scripts, such that two scripts are equivalent if they exhibit the same state for the same feature. In the case of identical feature states, the presumption could be that the shared state arose due to common inheritance. However, shared states can also arise due to homoplasies: borrowing (reticulate event) or random chance (autapomorphy).

A dendrogram is a common way to visualize the results of computational palaeographical analysis. There are different ways to represent this. One solution is when an internal node represents a palaeographical ancestor in a phylogenetic tree or network. Each taxon (usually script) is represented by a path (branch); the paths show

the different states as the writing system evolves. There is only one path between every pair of vertices. Another solution for a phylogenetic tree is one where the taxa are represented by nodes and their evolutionary relationships are represented by branches.

Script classification is carried out based on a taxon-feature matrix, where the rows usually correspond to the various scripts being analysed and the columns correspond to different features by which each script may be described; however, altering approaches are also possible depending on the goal of the actual investigation. Skelton (2007) used phylogenetic systematics for orthographical variations of the Linear B script, where taxa represented scribal hands and phylogenetic features represented variants of the same Linear B glyph. The phylogenetic tree produced by running the data matrix using parsimony as the optimality criterion is consistent with and clarifies what is known or hypothesized about the history of Linear B. Skelton demonstrated the usability of phylogenetic analysis to reconstruct the evolution of writing systems.

Wheeler and Whiteley (2014) criticised the use of basing analyses on proto-languages in historical linguistics, and their arguments apply equally to proto-scripts in the humanities-based palaeography. In such palaeography, classification of scripts is based on the comparison of graphemes, glyphs, and orthographical rules to identify regular correspondence features. From such features a proto-script is reconstructed, and it is posited as the evolutionary ancestor of the observed scripts. A proto-script is regarded as a real script once used by a population in a particular time and place. Differentially shared patterns of change from the proto-script among descendant scripts are used to determine subgroups within the family. However, variation is compounded by the inherently sporadic data: there are no records for several extinct scripts, which might have served to falsify proposed proto-script reconstructions. The concept of proto-scripts is used in this paper as a method to identify group scripts having a common unknown ancestor; however, no analysis is based on such theoretical scripts.

Constructing phylogenies based on the surviving inscriptions has some difficulties. Namely, several scripts were originally used with perishable carriers (papyrus, wood, etc.). Even for those scripts that are well represented, only certain parts of the inscriptions survived, which limits the range of features that can be examined. The record is usually just too incomplete in both a spatial and a temporal sense to be of much use. One possible approach is to construct a phylogeny based on the characteristics of surviving inscriptions. However, the available fossil record (corpus of surviving inscriptions) is so fragmentary that the phylogeny of the vast majority of taxa is unknown. Phenetic classification is possible for all groups. By contrast, cladistic analysis, based on branching sequences, requires historical inferences about the direction of evolution in a group of taxa (Lindberg 2012). Phenetics attempts to classify

taxa based on the concept of overall similarity, typically in morphology, without regard for their evolutionary relationships. Phenetic methods can be optimal when the distinctness of related taxa is important, and the data necessary for exploring the genetic relationships are missing. In phenetics, the more features on which the phenetic analysis is based, the better a given classification will be; every feature is of equal weight in creating natural taxa, and classifications are based on morphological similarity (Lindberg 2012). Subjectivity could be removed by examining as many features of the script as possible.

Decisions related to feature selection have the potential to impact a phylogenetic analysis, and these decisions also raise other issues, such as whether all features should be treated identically, or whether weighting schemes should be used to reflect the assumed reliability of the feature (Barbançon et al. 2013).

Consequently, in the case of limited information, it is futile to create an evolutionary tree, because there is no way to prove whether it is right or wrong. Instead, grouping taxa entirely on the basis of similarities is more efficient. As opposed to phylograms or cladograms, phenograms are only based on taxon similarities. In a phenogram, each branch point represents a step of increasing dissimilarity. In such case, the internal nodes of the graph do not represent ancestors but are introduced to represent the conflict between the different splits in the data analysis. The phenetic distance is the sum of the weights—represented as lengths—along the path between taxa. If discrete features are coded, the phenetic concept of homology is operationally identical to that used in cladistics. In phenetics, the homoplasy attending feature conflict is not reconciled (Wills 2001).

2.3 Machine-learning algorithms for phenetics

Phenetic analysis starts with the collection of raw measurement data on the chosen set of morphs, thus creating the taxon-feature matrix. Then a measure of dissimilarity is computed for each pair of taxa based on an appropriate metric. In the next step, a cluster analysis is performed to group taxa that are most similar. An index of average distance between each taxon could be calculated; then these distances are fitted into a hierarchical clustering pattern. It is difficult to decide which clustering algorithm should be used, and the methods do not all give the same answer. Therefore, following is an overview of some important algorithms used in the phenetic analysis, including the ordination, the clustering, the cluster validity indices, and the leaf ordering of the dendrograms to obtain a possible best cluster structure.

Clustering is an unsupervised learning (exploratory data analysis) method, which needs very little *a priori* knowledge. It is a useful technique for grouping data points such that points within a single group have similar characteristics, while points in different groups are dissimilar. Clustering is the task of categorizing objects having

	Features present in s_i	Features absent from s_i
Features present in s_j	f_{11} is the number of features present in both s_i and s_j	f_{01} is the number of features absent from s_i and present in s_j
Features absent from s_j	f_{10} is the number of features present in s_i and absent from s_j	f_{00} is the number of features absent from both s_i and s_j

Table 2: Parameters used in expressing the comparison of the features of scripts s_i and s_j

several attributes into different classes such that the objects belonging to the same class are similar, and those that are broken down into different classes are not. In the case of clustering, the problem is to group a given collection of unlabelled patterns into meaningful clusters. Labels are associated with clusters, but these category labels are data driven; that is, they are obtained solely from the data (Jain et al. 1999).

In clustering, the object is to place data points into the same cluster when they are similar enough according to some predefined metric. The predefined metric is one aspect that makes clustering a subjective process. In the case of computational palaeography, the features (variables) are the glyphs or orthographical rules, and the feature states (their values) are the presence or the absence of the glyphs or orthographical rules. Therefore, these variables are categorical with binary values. For comparing categorical data, the *Boolean* indicator variables are introduced. The formulae for the number of presence/absence feature states are written using the abbreviations in table 2.

The similarity of two objects (taxa, data points, in our case scripts) can be expressed by a metric. For categorical data, the *Jaccard* index (1) is widely applied, where M is the number of taxa. The Jaccard index is a statistic ordinarily applied to compare the similarity and diversity of the variables (features) of the examined objects, if the double absence (f_{00}) has no significance. This fits well with our dataset, since the clear majority of the features are glyphs, and the absence of a glyph in a script is not specific, since there are hundreds of glyphs that are absent from a certain script.

$$s_J(x_i, x_j) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}}, i, j \in \{1, \dots, M\} \tag{1}$$

The Jaccard index is not a metric; however, it can be converted to a metric distance, shown in (2).

$$d_J(x_i, x_j) = 1 - s_J(x_i, x_j) = \frac{f_{01} + f_{10}}{f_{11} + f_{10} + f_{01}}, i, j \in \{1, \dots, M\} \tag{2}$$

The square root of Jaccard distance is an Euclidean metric (Gower and Legendre 1986), given as (3).

$$d_{SRJ}(x_i, x_j) = \sqrt{d_J(x_i, x_j)}, i, j \in \{1, \dots, M\} \tag{3}$$

Another approach is to examine the object-variable (taxon-feature) matrix using a geometric representation: the objects (taxa) are points in a space spanned by variables (features) as axes of a scatter plot. Since the number of variables (dimension) is very large, it is necessary to replace the original large number of dimensions by a few artificial axes so as to represent the data structure as efficiently and faithfully as possible. This method is called *ordination* (Goodall 1954). One kind of ordination is *multidimensional scaling* (MDS), which can produce a dimension-reduction of objects from their dissimilarities. Where in the original high-dimensional space the variables of the objects are Boolean indicators, in the reduced dimensional space resulting from MDS the variables of the objects are quantitative.

A clustering approach can be taken not only in the original high-dimensional space but also in the reduced dimensional space. The *Squared Euclidean* difference is widely applied as a measure between quantitative data. Let $x_i = \left[x_{i\mathbf{1}}, x_{i\mathbf{2}}, \ldots, x_{iN}\right]$ and $x_j = \left[x_{j\mathbf{1}}, x_{j\mathbf{2}}, \ldots, x_{jN}\right]$ be two data points in the *N*-dimensional space of the data points. The Squared Euclidean difference is given as (4).

$$d_{SE}(x_i, x_j) = \sum_{k=\mathbf{1}}^{N} (x_{ik} - x_{jk})^2,\ i, j \in \{\mathbf{1}, \ldots, M\} \tag{4}$$

Clustering can be broken down into the following main steps. *Definition of object proximity:* as measured by a distance function defined on pairs of objects. *Clustering:* can be hard (crisp) or fuzzy. In crisp clustering, one object can belong to one and only one cluster. In fuzzy clustering, each object belongs to each cluster but with a varying degree of membership. *Cluster validation:* uses a specific criterion of optimality (Jain and Dubes 1988; Jain et al. 1999).

Jain et al. (1999) defined several types of clustering algorithms. *Hierarchical clustering:* These algorithms create clusters recursively by merging smaller partitions into larger ones or splitting larger clusters into smaller ones. These produce a nested series of clusters based on similarity. *Partitional clustering:* decomposes data sets into a set of disjointed clusters. *Density-based clustering:* creates clusters based on density functions. Its main advantage is to create arbitrary shaped clusters. *Grid-based clustering:* quantises the search space into a finite number of cells.

The diameter of a cluster can be defined in a number of ways. *Single linkage* (nearest neighbour) deals with the area where the two clusters are closest to each other. It emphasizes cluster separation: elongated point clouds are recognized, but clusters connected by intermediate objects cannot be detected. It is a hierarchical algorithm that can deal with arbitrary shapes, potentially at the expense of simple clusters. However, this tendency may also produce clusters that are chained. *Complete linkage* (farthest neighbour) deals with the whole area of the clusters; it is sensitive to outliers, and a single point far from the centre can greatly modify the clustering. It emphasizes cluster cohesion; the separation of clusters is not influential (Podani 2000). It produces

rightly bound or compact clusters (Baeza-Yates 1992). *UPGMA* (an unweighted pair group method of agglomeration, also called average linkage) merges in each iteration step the pair of clusters with the highest cohesion. In each grouping, the averages are calculated, and those groups with averages closest to each other are lumped together. It was developed for numerical taxonomy (Sokal and Michener 1958; Sokal and Sneath 1963). *WPGMA* (weighted pair group method, arithmetic average) uniformly weights all clusters independently of the number of their members. *Neighbour joining* (Saitou and Nei 1987) is based on the idea of parsimony; however, it does not attempt to obtain the shortest possible tree for a set of data. It operates on distance data, computes a transformation of the input matrix, and then computes the minimum distance of the pairs of objects. A weighted version of the method may also be used.

Ward's method minimizes the increase of the sum of squared deviations from the mean (Ward 1963). It optimizes the homogeneity of the clusters; it gives the most possibly homogenous clusters. In each step of the hierarchical clustering, Ward's method joins those two clusters where (5), the increase of the sum of squared deviations from the mean is minimal. In this case, M objects are clustered into a partition $C = \{C_1, \ldots, C_K\}$ of clusters, C_l and C_t are two different clusters ($l \neq t, 1 \leq l, t \leq K$, $C_l, C_t \in C$), $d_{Ward}\left(C_l, C_t\right)$ is the increase of the sum of squared deviations from the mean in the case of the fusion of C_l and C_t, and K is the actual number of clusters. Ward's method is appropriate for Euclidean distances, and it does not produce the clustering structure with the minimum error (Romesburg 2004, 129–135).

$$d_{Ward}(C_l, C_t) = \sum_{x_i, x_j \in C_l \cup C_t} d^2(x_i, x_j) - \left(\sum_{x_i, x_j \in C_l} d^2(x_i, x_j) + \sum_{x_i, x_j \in C_t} d^2(x_i, x_j) \right) \tag{5}$$

Another kind of clustering, called partitional methods, decomposes data sets into a disjointed cluster set. Such an algorithm is the *k-means*. It runs quickly but tends toward clusters with non-convex shapes. The k-means process minimizes the error E in (6)

$$E = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, \mu_i) \tag{6}$$

where μ_i is the center of cluster C_i, and K is the number of clusters. The number of iterations needed is unknown since standard k-means is not guaranteed to converge. Moreover, clustering produced by k-means is dependent on the starting points of the clusters.

Most clustering algorithms are very sensitive to their input parameters, and variations in the technique used can sometimes produce misleading results; verification through additional methods of dimensionality reduction analysis is essential, even though the ultimate objective of the research is classification (Podani 2000). Therefore, it is important to evaluate the result of the clustering process. Several clustering validity techniques and indices have been developed. The aim of cluster validity is

to find the partitioning that best fits the underlying data. Two measurement criteria have been proposed for evaluating and selecting an optimal cluster structure (Berry and Linoff 1996): *(i) Compactness:* The member of each cluster should be as close to each other as possible. A common measure of compactness is the variance. *(ii) Separation:* The clusters should be widely separated. The basis of comparison is the validity index. A validity index can provide a measure of the quality of the clustering on different partitions of a data set. It helps to determine the appropriate number of clusters present in a data set.

Dunn index is a cluster validity measure introduced by Dunn (1974) that maximizes inter-cluster distances while minimizing intra-cluster distances; it is a ratio of between-cluster and within-cluster separations. In other words, it is the ratio of the smallest distance between objects not in the same cluster to the largest intra-cluster distance, defined as (7)

$$D_i = \frac{\min\limits_{C_l, C_t \in C, l \neq t} \left[\min\limits_{x_i \in C_l, x_j \in C_t,} d(x_i, x_j) \right]}{\max\limits_{C_r \in C} \left[\max\limits_{x_i, x_j \in C_r,} d(x_i, x_j) \right]} \tag{7}$$

where M objects are clustered into a partition $C = \{C_1, \ldots, C_K\}$ of clusters, $d\left(x_i, x_j\right)$ is the distance between objects $x_i \in C_l, x_j \in C_t, l \neq t$, and $\min\limits_{x_i \in C_l, x_j \in C_t,} d\left(x_i, x_j\right)$ is an intercluster distance metric between clusters $C_l, C_t \in C, l \neq t$. High Dunn index means that the diameter of the clusters is small and the distance between clusters is large; therefore, the clusters are compact and separated. This measurement serves as a measure to find the right number of clusters in a data set, where the maximum value of the index represents the right partitioning given the index. Its disadvantage is that it is sensitive to noise, because the maximum cluster diameter can be large in a noisy environment.

Silhouette index is another approach to measure how similar a given object is to objects in its own cluster, as compared to objects in other clusters. Silhouette is higher when clusters are dense, well separated, or convex, and a zero value indicates overlapping clusters. It provides a graphical representation of how well each object lies within its cluster (Kaufman and Rousseeuw 1990). The S_i Silhouette index for the object x_i is defined as (8)

$$S_i = \frac{b_i - a_i}{max\,(a_i, b_i)} \tag{8}$$

where a_i is the average distance from the object x_i to the other objects in the same cluster, and b_i is the minimum average distance from the object x_i to objects in a different cluster, minimized over clusters. The range of Silhouette index is $[-1, 1]$. A high Silhouette value indicates that the object x_i is well matched to its own cluster, and poorly matched to neighbouring clusters. If most objects have a high Silhouette

value, then the clustering solution is appropriate. The Silhouette index can be used with any distance metric.

Trees that result from cluster analysis are typically presented with their leaves in an undefined order. However, the distance of these leaves in the dendrogram could refine the cluster structure. Therefore, it is important to maximize the sum of the similarity of adjacent objects in the dendrogram. In the hierarchical clustering investigations, an optimal *leaf ordering* for hierarchical binary cluster tree (Bar-Joseph et al. 2001) was applied.

2.4 Notation and palaeographical sources

Different "runiform" or "Runic"-type scripts were used in largely different places in the Eurasian Steppe and in the Carpathian Basin; their surviving inscriptions are mainly from the 1st millennium AD. Their possible relationship has not been proved or widely accepted. Furthermore, a lot of inscriptions of the Eurasian Steppe have not been deciphered yet. However, many authors have previously demonstrated the similarities of the scripts used in some of these inscriptions (Nagy 1895; Sebestyén 1915, 143–160; Németh 1917–1920; Ligeti 1925; Kyzlasov 1994; Vasil'ev 1994; Vékony 1987b, among others).

Unlike in earlier attempts to decipher them, a single acknowledged scholar, G. Vékony (late Assoc. Prof. in the Eötvös Loránd University, Budapest, 1944–2004), provided a comprehensive decipherment for several of the inscriptions from the Carpathian Basin to Middle Asia. Therefore, his decipherment, including the determined sound values of the signs, was used in the phenetic analysis in this paper. Vékony published his results in several publications (1981; 1985a; 1985b; 1987a; 1987b; 1992a; 1992b; 1992c; 1993; 1996; 1999a; 1999b; 2004), mostly in Hungarian. Since 2008, the author of this article has systematically consulted with acknowledged scholars (linguists, archaeologists, historians), who validated and improved the readings of Vékony. The results of these collaborations are published in English (Hosszú 2012), and in Hungarian (Hosszú 2013; Hosszú and Zelliger 2013; 2014a; 2014b). It should be emphasized that computational palaeography uses the results of humanities-based palaeography. The author utilised these results as accurately as possible.

The very close similarities between some of the scripts of the Eurasian Steppe are demonstrated as an application of the phenetic method; however, there is no category name for these scripts. In the literature, mostly the terms "runiform" or "Runic" are used, which are largely inappropriate, since these scripts are fundamentally different from the Runic script and its various versions (older fuþark, Anglo-Saxon runes, younger/Danish fuþark, short-twig runes, etc.). In Hungarian scientific literature, these scripts are have, for the last century, usually been called "rovásírások" 'Rovash scripts' (e.g., Sebestyén 1909; 1915). Therefore, the author collectively calls these

scripts "Rovash." It is noteworthy that modifying the name of a script based on the research results is not unknown. For instance, the Anatolian Hieroglyphic script was earlier denoted as Luwian hieroglyphic, and even earlier Hittite Hieroglyphic (Payne 2010, 2; Yakubovich 2015a, 5). Another example is the Cypro-Greek script, which name was proposed by Egetmeyer in 2010 to replace the traditionally used "Cypriot Syllabary."

Table 3 presents some abbreviations and symbols used throughout the paper.

/	The alternative hypotheses are separated by a slash.
/ /	Double slashes denote phonemic transcription (denoting phonemes), phonemic representation of grapheme.
?	Question mark denotes the non-consensual transcription or phonetic value.
[ ]	Square brackets denote phonetic transcription (denoting allophones) using IPA (International Phonetics Association) symbols. The square bracket denotes the optional texts, too.
< >	Angled brackets are used for denoting transliteration value. In transliteration, the case that a consonant used before or after a sound is denoted by writing the transliteration value of that sound in superscript, e.g. <$^{\ddot{w}}$k$^{\ddot{w}}$>.
<A>	Transliteration value of Rovash graphemes with /a, ä/ phonemes.
<W>	Transliteration value of Rovash graphemes with /o, u/ phonemes.
<Ẅ>	Transliteration value of Rovash graphemes with /ö, ü/ phonemes.
↔	A part of a sound continuum; e.g., /c ↔ ʦ/ is a part of the continuum [k] > [kj] > [c] > [cç] > [cɕ] > [tɕ] > [ʧ] > [ʦ] > [s], or /ɟ ↔ ʣ/ is a part of the continuum [gj] > [ɟ] > [ɟj] > [ɟʒ] > [ʤ] > [ʤ] > [ʣ] > [z] (based on Valério 2016, 217, 256, 259).
AH	Anatolian hieroglyphic (Luwian / Luvian / Anatolian Hieroglyphic / hieroglyphs / syllabary / syllabic) script
AGA	Anatolian-Greek alphabetic scripts
C	Consonant
CBR	Carpathian Basin Rovash (Nagyszentmiklós, Tisza) script
CGk	Cypro-Greek (Valério 2016) script
CM	Cypro-Minoan script
CT	Characteristic Transformation (in a topological layer of the layered grapheme model)
Cypro-Greek	Cypro-Greek syllabary (Cypriot syllabic, Cypro-Syllabic, Classical Cypriot syllabary, Linear C). The term *syllabaire chypro-grec* was introduced by Egetmeyer (2010) and supported by Valério (2014).
dextrograde	Left-to-right writing (direction of writing)
E. Cyrillic	Early Cyrillic script
I. Aramaic	Imperial Aramaic (*Reichsaramäisch*, Official/Standard Aramaic) script
Lin. A	Linear A script
Lin. B	Linear B script

Madhabic	Instead of the earlier Minaic, Macdonald (2000, 68) recommended use of Madhabic.
NE-Iberian	Northeastern Iberian (Levantine Iberian) script
Old Aramaic	*Altaramäish,* Ancient/Early Aramaic script
ONA	Oasis North Arabian is a script group, its members: Dumaitic, Taymanitic, Dadanitic, and Dispersed North Arabian (Macdonald 2004, 490).
P.-Campanian	Proto-Campanian (*Protocampano, Paleoitalico, Nucerino alphabet*) script
P.-Canaanite	Proto-Canaanite script
P.-Hebrew	Palaeo-Hebrew script
P.-Hispanic	Palaeo-Hispanic (Palaeohispanic) script family
P.-Sinaitic	Proto-Sinaitic script
P.-Umbrian	Palaeo-Umbrian script
Proto-Rovash	The supposed common ancestor of the Rovash scripts (TR, SR, CBR, SHR), as hypothesized by the author.
S. Picene	South Picene script
S. Semitic	South Semitic script family
SE-Iberian	Southeastern Iberian (Meridional Iberian) script
SFG	Similarity Features Group
SHR	Székely-Hungarian Rovash (Székely, *Sekler,* [Old] Hungarian) script
sinistrograde	Right-to-left writing (direction of writing)
SR	Steppean Rovash (Khazarian Rovash, Don-Kuban-South-Yenissei-Ačïqtaš-Isfar, East European Runic Script). Note that the meanings of these scripts (the sets of inscriptions classified to each of them) partly differ from each other.
stiktogram	Punctuation mark (Karnava 1999, 37).
SW	Southwestern (Southwest, South Lusitanian, Tartessian, Bastulo-Tartessian, Southern Portugal) script
syllabogram	A grapheme that represents a syllable.
Th.	Thamudic is the tentative name of Ancient North Arabian scripts, which differs from the well-defined ONA, Safaitic, or Hismaic scripts. The Thamudic scripts are the following: Th. B, Th. C, Th. D, and Southern Th. (Macdonald 2004, 492).
TR	Turkic Rovash (Orkhon-Yenissei-Talas, [East/Old] Turkic runiform / Runic / "Runic") script. In this paper, the term *Turkic Rovash* is used instead of the more common *Turkic runic,* since this script is grouped together with the other Rovash scripts to avoid confusion with the fundamentally different Runic script.
V	Vowel

Table 3: Abbreviations, alternative names, and symbols

In representing graphemes, if the phonemic transcription (e.g., /b/) is obviously based on the transcription value (e.g., <b>), the phonemic transcription is usually not

denoted, to simplify the description. Note that in some sources, only the transliteration values are available. Moreover, there are several differences between the sound values of the same grapheme in different sources. Therefore, the presented computational palaeographical analysis can further be made more accurate depending on the new results of the palaeography of the humanities.

The graphemes are identified by the script name and the transliteration value (e.g., AH <kar>, CGk <ko>, CM <ko?>, Greek <α>, NE-Iberian <ga/ka>) or the script name and the grapheme name (e.g., AH *315, CM 15). If the grapheme name includes the abbreviation of the script name (e.g., CM 15), the script name can be omitted. Usually, one or more typical glyphs are also included in the grapheme identification (e.g.: AH 𐤟, 𐤊, 𐤟 *315 <kar>; NE-Iberian 𐬀 <ga/ka>).

In the case of sequencing graphemes, usually the graphemes are separated by a semicolon (;). However, the repeated identical script names, glyphs, or transliteration values are omitted in order to save space. In the case of an omission, the glyphs are separated by a comma (,), e.g., "Carian 𐊠, Lydian 𐤠 <a>" is written instead of "Carian 𐊠 <a>; Lydian 𐤠 <a>." Another example: "Taymanitic, Hasaitic 𐪩 <y>" is written instead of "Taymanitic 𐪩 <y>; Hasaitic 𐪩 <y>".

The use period of the examined scripts (table 4) are mostly estimations due to the inaccuracies of dating archaeological relics, and since if in a certain period a script was written on perishable materials, no relic survived. In table 4, the use periods of closely related scripts are given collectively. The grouping of scripts is based on historical and phenetic features, and not on proved genealogical relationships. If the sound value of a grapheme of any script has not been proved, a question mark (?) denotes this fact, and such grapheme is omitted from the numerical analysis.

The Lin. A and CM scripts are still undeciphered; therefore the author omitted all of their graphemes from the numerical analysis. However, in the case of several signs, there is a consensus about their probable sound values (Valério 2016); therefore, several Lin. A or CM graphemes were included in the SFGs for information purposes.

The sources of the palaeographical data of graphemes and scripts are generally not detailed in table 10 due to the very large number of used glyphs. The author used the glyphs and other palaeographical statements from the sources table 5.

3 Method

3.1 The concept of the developed method

In computational palaeography, the variability of a grapheme could easily result in identical glyphs of unrelated graphemes; therefore, identical feature states could appear without any genealogical relationship (homoplasy). Consequently and significantly, the identity of a computational palaeographical feature state is generally

Groups	Estimated period of use of scripts
Aegean	*CGk*: 11th–2nd c. BC; *CM*: 17th/16th–11th c. BC; *Lin. A*: 18th–14th c. BC; *Lin. B*: 15th–13th c. BC
AH	*AH:* 17th–7th c. BC
AGA	*Carian:* 7th–3rd c. BC, *Greek:* 8th c. BC –, *Lemnian:* 6th c. BC; *Lycian:* 5th–4th c. BC, *Lydian:* 8th–3rd c. BC, *Phrygian* (archaic period only): 8th–4th c. BC, *Sidetic:* 5th–2nd c. BC
Ancient Italic	*Camunic, Elymian, Etruscan, Faliscan, Gallo-Etruscan, Gallo-Greek, Gallo-Latin, Latin, Lepontic, Messapic, Oscan, P.-Umbrian, P.-Campanian, Raetic, S. Picene, Umbrian, Venetic:* 8th c. BC – 1st c. AD
Ancient Semitic & Canaanite	*P.-Sinaitic & P.-Canaanite:* 19th (?) – 11th (?) c. BC; *Old Aramaic:* 925–700 BC; *Phoenician, P.-Hebrew, Samaritan:* 15th c. BC – 2nd c. AD
Aramaic & Persian	*Arabic:* 6th c. AD –; *Hatran:* 1st c. BC – 3rd c. AD; *Hebrew:* 3rd c. BC –; *I. Aramaic:* 700–200 BC, *Middle Persian:* 3rd–7th c. AD; *Nabataean* 2nd c. BC – 4th c. AD; *Palmyrene* 1st c. BC – 3rd c. AD; *Parthian:* 2nd c. BC – 3rd c. AD; *Sogdian:* 3rd–13th c. AD; *Syriac:* 1st c. AD –
Libyco-Berber	*Libyco-Berber:* 8th/7th c. BC – 7th c. AD
P.-Hispanic	*Celtiberian, NE-Iberian, SE-Iberian, SW:* 8th–1st c. BC
Rovash	*CBR:* 6th–11th/12th c. AD; *SHR:* 10th c. AD –; *SR:* 8th–12th c. AD; *TR*: 7th–10th c. AD
Runic	*Runic:* 2nd c. BC –
S. Semitic	*Ancient North Arabian (Taymanitic, Dadanitic, Dumaitic, Dispersed ONA, Safaitic, Hismaitic, Th. B, Th. C, Th. D, Southern Th.):* 8th c. BC – 4th c. AD; *Ancient South Arabian (Sabaic, Madhabic, Hasaitic):* 11th c. BC – 6th c. AD; *Geʿez abjad:* 8th c. BC – 4th c. AD
Slavic	*E. Cyrillic:* 10th c. AD –; *Glagolitic:* 9th c. AD –

Table 4: The groups of scripts and the estimated use periods of scripts based on surviving inscriptions

less well determined than in biology (e.g., gene sequence). In order to obtain identical feature states after filtering out the homoplasies, the linguistic, historical and geographical circumstances must be taken into account along with the topological similarities of the glyphs (table 9).

It is difficult to directly determine a script's genealogy in part due to the long examined period (generally from 2nd millennium BC to 1st millennium AD), and during this time frame scripts may have influenced each other on multiple occasions. Moreover, presently unknown scripts and orthographies may have existed that may also have influenced the examined scripts. However, by narrowing the focus of study to individual graphemes, connections might be determined. A slightly similar concept

Adiego (2007a; 2007b; 2007c; 2007d; 2007e; 2015), Anders (2012), Bakkum (2009), Benkő (1996a; 1996b), Beyer (1998), Bordreuil (2005), Brixhe and Lejeune (1984), Colless (2010), Correia (1996), Cross (1989), Daniels and Bright (1996), Davies and Olivier (2012), Davis (2010), Doblhofer (1962), Erdal (1993), Eska (2008), Farrujia de la Rosa et al. (2010), Faulmann (1880), Ferrer i Jané (2005; 2013; 2014), Gabain (1941), Garbini (1979), Gibson (1975), Grimme (1923), Hampel (1884), Hawkins (1986; 2000; 2010), Healey (1990), Hempl (1899), Hesperia (2005), Hoffmann (1987; 2011), Hosszú (2012; 2013), Hosszú and Zelliger (2013; 2014a; 2014b), Jeffery (1961), Jensen (1969), Kairzhanov (2014), Kalinka (1901 *apud* Adiego 2015), Kara (1996), Karali (2007), Kenyon (1899), King (1992), Konkobaev et al. (2015), Kononov (1980), Krings (1995), Kyzlasov (1994), LBI, Lemaire and Sass (2013), Looijenga (1997), Macdonald (2004; 2005; 2015), Marchesini (2009; 2012; 2014), MacKenzie (1971), Masson (1976; 1978), Mees (2006), Melchert (2004; 2008a; 2008b, 2008c), Miller (1994), MNAMON, Morandi (1982; 2004), NLR, Nollé (2001), O'Connor (1996), Olivier and Vandenabeele (2007), Olivier (2007-2008), Payne (2010), PROEL, Rilly and de Voogt (2012), Rodríguez Ramos (2004), Rogers (1999), Rollston (2008), Róna-Tas (1987), Rosenthal et al. (1986-2011), Röllig (1995), Sándor (1991), Sass (1988), Sebestyén (1915), Sims-Williams and Grenet (2007), Skjærvø (1996), Sprengling (1931), Swiggers (1996), Swiggers and Jenniges (1996), Taylor (1883), Tekin (2003), Thelegdi (1994/1598), Thompson (1912), Thomsen (1893), Tzanavari and Christidis (1995), Urbanová (2003), Valério (2008; 2013; 2016), Vékony (1985a; 1987a; 1992a; 1999a; 1999b; 2004), Wallace (2007), Weeden (2014), Woodard (1997; 2014), Woudhuizen (1982–1983; 1984–1985a; 1984-1985b), Yakubovich (2015), Young (1969), Younger (2000; 2003–2012).

Table 5: The sources of the palaeographical data

was proposed by Bernal (1990), who traced "isographs" of each grapheme instead of whole scripts.

The comparison of the glyphs of different scripts is supported by Boisson's stability principle, i.e. graphemes representing a sound existing in the acceptor language are adopted with their original glyph and sound value (Boisson 1994, 225 *apud* Adiego 2007e, 2).

Macdonald (2015, 10–12, 28–29) differentiated between literate and non-literate societies. He only considered a society literate if the written word was essential to its day-to-day functions. It is not necessary for the majority in the society to be able to read and write, but if a society had a written script and members who could read and write, but the skill was not used in everyday life, such a society is considered non-literate. According to Macdonald, most nomadic societies were non-literate, or mostly non-literate. An inscription in an illiterate society does not serve practical reasons, such as the majority of the Safaitic inscriptions on the boulders scattered in the desert. The reader is less important, so reader requirements do not affect the development of the script. Typically, writing is continuous with no word-dividers. Occasionally decorative variants were created in particular inscriptions, but had no consequences on the stability of the script itself.

Macdonald is aware that many societies cannot be strictly confined to a single category (literate or non-literate) but are in transition from one to the other. Even

the scripts used exclusively by nomadic societies show development (i.e. new glyph variants becoming widespread), meaning that the written word must have had readers who were able to select the new alternative of a glyph. Applied to the Rovash scripts (used in the Altai Mountains, the Eurasian Steppe, and the Carpathian Basin; the majority of the surviving inscriptions are read in Turkic and Hungarian; see some examples of Rovash inscriptions in the Appendix), they must have been used largely in non-literate societies, as widespread modifications are largely absent. Thus Macdonald's model gives the basis to compare Mediterranean glyphs from the 1st millennium BC with Rovash scripts, which are attested only after the 6th c. AD (table 4). The only extant Rovash script, Székely-Hungarian Rovash (SHR), was used in the relatively isolated community of the Székelys (living in the mountains of Transylvania) up to the 17th c. AD. The Székelys used the Hungarian orthography of the Latin script for day-to-day functions; SHR was used as an unofficial writing system, and knowledge of SHR was passed almost exclusively from father to son. Consequently, in the case of the Székely-Hungarian Rovash (SHR) script, inscriptions made up to the 17th c. AD can be used for the present analysis.

3.2 Conversion of the palaeographical data into similarity features groups

In the analysis, the objects (taxa) are the scripts, and the features of a script are their graphemes and orthographic rules. The main properties of graphemes are the glyphs, especially their shapes. Other features of the scripts are their orthographical properties. Since these features are categorical variables, they are transcoded to Boolean indicators with the value being 1 if the feature is present in a script and 0 otherwise. The input data are given by the matrix $X(S, F_x)$ where S is the vector of scripts (objects, taxa) and F_x is the vector of features (presence of glyphs or orthographic rules). In this matrix a total of 66 scripts have been recorded with over 186 features. An illustrative example of the $X(S, F_x)$ matrix is given below as equation (9), which is generated from the similarity features groups (SFGs), where the features are the presence of glyphs or orthographical rules in a script (table 6).

$$X(S, F_x) = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \tag{9}$$

Automatic comparison of glyphs is an inherently difficult problem because *(i)* related glyphs may be realised vastly differently in inscriptions due to differing

	Celtiberian	Elymian	Linear B	NE-Iberian	Raetic	SE-Iberian	SR	SW	TR
𐌀 <a> (SFG-10)	0	1	0	0	1	0	0	0	0
𐌁, 𐌂 <b, bo, bu> (SFG-123)	1	0	0	1	0	1	1	1	0
𐌄, 𐌅, 𐌊 <ge/ke, k^2> (SFG-94)	1	0	0	1	0	0	1	0	1
𐌍, 𐌋 <ga/ka, g> (SFG-89)	1	0	0	1	0	1	0	0	0
𐌇, 𐌈 <pa_3, p^u> (SFG-150)	0	0	1	0	0	0	0	1	0
Boustrophedon in some relics (SFG-182)	0	0	0	1	0	0	0	1	1

Table 6: An illustrative example of $X^T(S, F_x)$ the transpose of data matrix (abbreviations in table 3)

calligraphic requirements, and *(ii)* unrelated glyphs may take very similar shapes when written. To describe the written variations of a glyph, the multilayer grapheme model has been developed, where each grapheme's visual identity has been determined and it was further assumed that in a script at any given period the glyphs may only diverge in so far as the common visual identity remains intact (table 8).

To handle the visual differences between glyph variants representing one grapheme, a typical set of transformations has been developed (table 6) that can describe how the shape of one glyph variant transformed into another. Using these transformations, it is easier to decide if certain symbols of the inscriptions are glyph variants of each other or not; however, it has to be noted that this is not a sufficient condition. Even with taking into account all available data, deciding on the relationship between two symbols in the same script or even in different scripts is uncertain. The more palaeographical data is taken into account, the less the uncertainty. The known glyph variants of one or more scripts are collected into a similarity features group (SFG). As the available palaeographical data (shapes of symbols in inscriptions, age of the inscriptions, published set of inscriptions, sound values of the graphemes, orthographical properties of the scripts, etc.) increases, the SFGs are split or restructured to fit with the new data. The more palaeographical data are analysed, the easier it is to build more realistic SFGs, and as a consequence, the size of the $X(S, F_x)$ matrix is usually increasing.

Macdonald (2015, 18–22, 24–26, 34–35, 40) criticised *comparative palaeography,* in which grapheme chains were stated as sequences of development. Among others, he cited the theories of Lidzbarski (1902, 122) and Praetorius (1904, 717–718). Lidzbarski proposed the genealogical relationship between the Phoenician 𐤀 <'>, Dadanitic 𐪀 <'>, and Safaitic 𐪉 <'>; Praetorius improved upon Lidzbarski's theory and suggested a

sequence between Phoenician 𐤀, 𐤊 <'>, Safaitic 𐪊 <'>, and Dadanitic 𐩠 <'>. Macdonald (2015, 34) criticised Lidzbarski's opinion that there was a tendency towards modifying the irregular shapes of the North Semitic graphemes into symmetrical glyphs in both the South Semitic and the Greek scripts. Macdonald (2015, 35, 41) claimed that there is no evidence for any progressive development of the known South Semitic glyphs; and he presented examples of homoplasies, when a glyph in one script can develop a form similar to that of its equivalent in a different script.

Based on Macdonald's arguments, the present research is restricted to the collection of SFGs, and not the complete genealogical sequences, in order to minimize possible errors. However, there was evolution in glyphs; consequently, applying the results of the type of palaeography found in the humanities and in phylogenetics, a genealogical model must be achieved, too. Moreover, in the case of certain SFGs, there are glyphs that are obviously relatives; however, their shapes are slightly different. For instance, the Rovash 𐵇 <χ> (SFG-101) and 𐵆 <k^5> are cognates; however, the characteristic transformation *shortening lines* (CT-8, in table 7) can be applied to transform one to the other. It is worth noting that the presence of a characteristic transformation between two presumably cognate glyphs is only a supposition, and more palaeographical data could falsify or support its presence.

The objective in these examinations is to use the actual realised glyphs in inscriptions and not idealised glyphs. However, the conventions of the literature, which forms the basis of the $X(S, F_x)$ matrix, differ widely—many scientific articles publish the inscription only using idealised glyphs, while others publish faithful drawings of the inscriptions.

3.3 The developed exploratory data analysis

In order to explore the similarities of historical scripts (objects, taxa), a composite phenetic analysis method was developed, presented as the flow chart in figure 1, where X is the taxon-feature data matrix, Y is the taxon-feature data matrix transformed into 2-dimensional or 3-dimensional space, F_{MDS} is the vector of transformed features of the taxa, C is the matrix of cluster configuration, Z is a tree of hierarchical clusters, I is the vector of the cluster identifiers, and K is the actual number of clusters.

One result of the present research is the multilayer model of the graphemes, which was developed for modelling the grapheme in computational palaeography. The developed grapheme model is composed of four logical layers from bottom to top, namely the Topology, Visual Identity, Phonetic, and Semantic Layers. In the Topology Layer, a single glyph is described by a complete set of geometrical attributes. The Visual Identity Layer focuses on determining the possible unique identity of a writing symbol based on the human visual perspective in identifying an object. In this layer, the various glyphs of a single grapheme share some topological attributes in common.

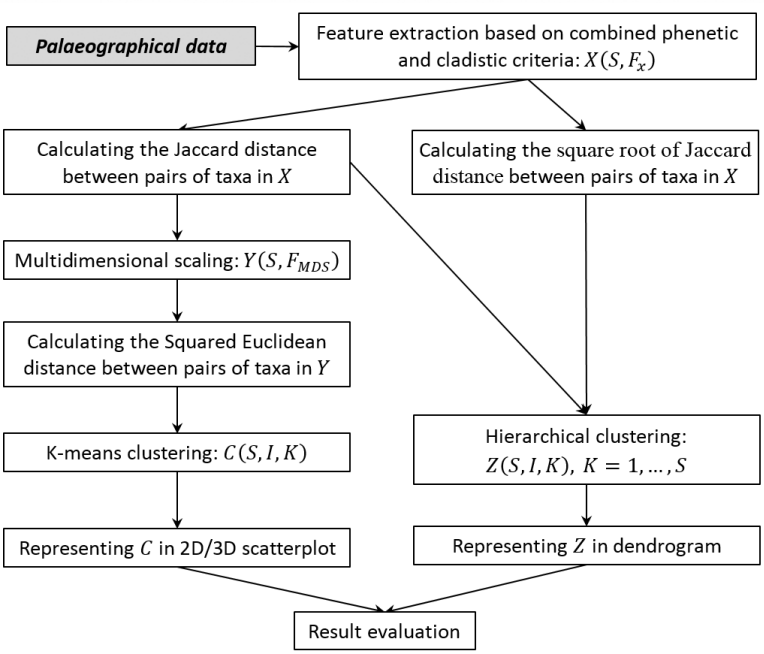


Figure 1: The flow chart of the developed method

The Phonetic Layer gives the sound values associated with the grapheme, and the Semantic Layer takes into account the context of the usage of the grapheme in the surviving and deciphered inscriptions (Pardede et al. 2016).

The Topology Layer of cognate, albeit slightly different glyphs can be transformed into each other by topological transformations called *characteristic transformations*. The characteristic transformation (CT) usually does not change the visual identity of the original glyph. Some examples of these transformations are listed in table 7, where there are references to SFGs in table 10.

Table 8 presents the four-layer grapheme model for the NE-Iberian <be> grapheme. This model helps to differentiate between the less important glyph variants and the significant altering graphemes. The CTs in table 7 are ideal geometric transformations in the topology layer of the grapheme model in table 8; however, the actual realization in the glyphs' evolution is unique in each case. It is noteworthy that on one hand, two glyphs are not necessarily relatives even if their shapes are identical or the difference can be covered by a CT; and on the other hand, the differences between cognate glyphs can usually be covered by CTs.

CT-1: *Bending or straightening,* e.g.: TR , , <t[1]> (SFG-29). Carian , , , <d> (SFG-30). SHR , <e> (SFG-68). NE-Iberian , <ge/ke> (SFG-95). Libyco-Berber , <R> (SFG-160). Etruscan , <fh> (SFG-166).

CT-2: *Closer-shape forming or vice versa,* e.g.: In the following two cases, Macdonald reconstructed the sequence of glyph development: Dadanitic > > > <'> (SFG-3), (Macdonald 2010, 12–14) and Dadanitic > > > <s[1]> (SFG-142) (Macdonald 2010, 13–14). Other example: TR , , , , <t[1]> (SFG-29); TR , CBR , SHR <W> (SFG-44).

CT-3: *Connecting,* e.g.: TR , <g[1]/[1]γ> (SFG-103). CM CM 56 (SFG-133), TR <n[2]>.

CT-4: *Cursivizing,* e.g.: Sabaic (early zabūr) , (middle zabūr) < ṯ > (SFG-166).

CT-5: *Increasing or decreasing the number of repeating lines or curves,* e.g.: SHR , <ď> (SFG-52). P.-Campanian , <s> (SFG-58). SHR , <z> (SFG-59). P.-Sinaitic , <ḫ> (SFG-68). S. Semitic , <ḍ> (SFG-70). AH , *315 <kar> (SFG-90). CBR , , <z> (SFG-92). Runic , <b> (SFG-119). SW , , <s> (SFG-168).

CT-6: *Line insertion or deletion,* e.g.: NE-Iberian , , , <o> (SFG-66); NE-Iberian , , <be> (SFG-116).

CT-7: *Loop opening or vice versa,* e.g.: Dadanitic , <ḏ> (SFG-55); NE-Iberian , <be> (SFG-116).

CT-8: *Shortening of lines,* e.g.: SR <k[5]> /q/, SHR <χ> /χ/ (SFG-101), the differences in the sound values are linguistically justifiable (Vékony 2004, 108–109). Sabaic zabūr (early) , (middle) <z> (SFG-53).

CT-9: *Straight to curve or vice versa,* e.g.: SR <k[5]>, CBR <q> (SFG-101). NE-Iberian , <be> (SFG-116). Safaitic , , , , , , Hasaitic <ṯ> (SFG-166) (Macdonald 2005, 82; 2015, 37). Greek (Corinth) , <ε> (Swiggers 1996, 264). Greek (cursive, 6[th] c. AD) , , <β> /b/ (Thompson 1912).

CT-10: *Turning ±90°,* e.g.: TR <b[2]> and <m> (SFG-116). It is a typical Anatolian feature (Woudhuizen 1984–1985a, 92). Carian , <λ>. Carian , <ś> (SFG-170). AH , , (Payne 2010, 14, 79), (Anders 2012) *412 <ru>.

CT-11: *Vertical mirroring,* e.g.: Old Aramaic , Greek <b> (SFG-13); NE-Iberian , <be> (SFG-116).

Table 7: Examples of characteristic transformations

4 Realization

The main goal of this paper is to demonstrate that the developed exploratory data analysis algorithm is applicable to processing palaeographical datasets and evaluating their statistical modelling. The realization of the method is presented below.

Layers	Example
Semantic	It was used in a semi-syllabic script in the period of 5th–1st c. BC in Northeastern Iberia (today Spain) and the Roman province Gallia Narbonensis (today France).
Phonetic	/be/
Visual Identity	A loop like shape with at least two legs up or down.
Topology	Glyph variants: 𐊘, 𐊙, 𐊚, 𐊛, 𐊜, 𐊝, 𐊞, 𐊟, 𐊠, 𐊡, 𐊢, 𐊣, 𐊤, 𐊥, 𐊦, 𐊧, 𐊨, 𐊩, 𐊪. The applied topological transformations (citing CTs in table 7) are line insertion or deletion (CT-6), loop opening or vice versa (CT-7), straight to curve or vice versa (CT-9), and vertical mirroring (CT-11).

Table 8: Example of the four-layer grapheme model for the NE-Iberian <be> (SFG-116 in table 10)

C-1: The sound values of the graphemes are identical, or the difference is linguistically justified by acknowledged scholars.
C-2: Typologically the examined glyphs are identical or their difference is reasonable.
C-3: The historical and geographical facts prove or at least do not rule out the relationship between the scripts of the examined graphemes.
C-4: In the case of phonetically or topological differences, such SFG structure is chosen in which the supposed number of evolutionary changes is minimal.

Table 9: Criteria for constructing similarity features groups (SFGs)

4.1 Feature extraction based on similarity of glyphs and orthographical rules

In collecting the members of each SFG, the conjunctive criteria C-1 and C-2 (table 9) for the assumption of borrowing were considered. If both C-1 and C-2 are met, the appropriate glyphs are taken to be a member of a same SFG. This procedure is a phenetic analysis of the scripts. However, there are two problems: First, in the cases of several glyphs, based on the conditions above, a glyph could be classified into more than one SFG. Second, some glyphs that fulfill the criteria C-1 and C-2 could be homoplasies. Therefore, two further criteria were added in the analysis, C-3 and C-4 (table 9). The criterion C-3 helps to identify the homoplasies and to select such glyphs into separate SFGs. By using the criterion C-4, the developed combined method is governed by the *lex parsimoniae* (Ockham's razor) as is usual in cladistics.

When constructing the SFGs, the scientific literature has been taken into account, especially the dissertation of Valério (2016). This article thus primarily attempts to

show the usefulness of phenetic modelling in developing the SFGs; phonetic and topological similarity is only used as a rough guide. In more specific palaeographical analyses, the method may be made more precise. While the SFGs used in this article are novel concepts and do not appear in the general literature in this form, similar approaches have been used, e.g., groups of presumably cognate graphemes in Valério (2016).

The explored SFGs are presented in table 10; the SFGs are sequence numbered. In each SFG, the topology of the glyphs or the orthographical rules is similar or identical. It is noteworthy that in the case of each grapheme, mainly those glyphs are listed that best fit in the appropriate SFG. The number of possible SFGs could be several hundred, but the set of SFGs is limited to those 186 that are the most significant. In the performed numerical analysis, only the presented SFGs are used. Naturally, more SFGs means more accurate results.

Analysing phonetic changes belongs to the palaeography of the humanities, and is outside the scope of computational palaeography. For example, SFG-91 is based on the combination of palaeographical and phonetic analysis (Valério 2016, 253–256). Similarly, SFG-92 is only a proposal, since it lacks palaeographical and linguistic evidence.

In each cell of table 10, first the members of the actual SFG are listed with their script names in italics. Following them are comments, which could contain further graphemes. However, the graphemes occurring only in comments are not included in the SFGs or any numerical analysis.

In general, the sources of individual glyphs in this article are not individually cited, since those may be found in the reference material (table 5). Furthermore, the name and age of the inscription in which a glyph is found is omitted for brevity, except in critical cases where these data are important for the analysis. In constructing the SFGs, in addition to the properties of the graphemes (glyphs, transliteration values, grapheme name, sound values), proposed relationships of different graphemes were obtained from palaeography publications from the humanities. Due to the large number of SFGs, in most cases, the author of the present paper could not detail all data from the scientific sources; therefore, publications used in constructing the SFGs are collected in table 5. Note that the goal of this paper is to present a developed phylogenetic procedure optimized for palaeographical data, and not to offer detailed palaeographical analyses (which is the task of humanities-type palaeography).

SFG-1: *P.-Sinaitic* 𐤀, 𐤄; 𐤉, *P.-Canaanite* 𐤀, 𐤃, *Phoenician* (Nora, ca. 900 BC) 𐤀, (Kilamuwa, ca. 825 BC) 𐤀, (Cyprus, ca. 880 BC) 𐤀, (Limassol, ca. 750 BC) 𐤀, 𐤀, 𐤀 <’>; *Phrygian* A, *Greek* 𐌀, A, *Lemnian* A, *Carian* A, A, *Lydian* A, *Elymian* A, *S. Picene* A, *Etruscan* A, A, *Raetic* A, A, *Faliscan* A, *Venetic* A, *Messapic* A, *Lepontic* A, *P.-Umbrian* A, *Umbrian* A, A, *Oscan* A, *Latin* A, A, *Gallo-Greek* A, *SW* A, A, *E. Cyrillic* A, A <a>

SFG-2: *Greek (medieval cursive)* 𐌰 (Faulmann 1880, 171) <α>; *Glagolitic* ⰰ *azъ* <a>

SFG-3: *Sabaic* 𐩱, *Dispersed ONA, Dumaitic, Taymanitic, Th. B, Th. D* 𐩱, *Dadanitic* 𐩱, 𐩦, 𐩦, 𐩧 <’> /ʔ/. *(i)* Glyph evolution CT-2.

SFG-4: *Phoenician* (al-Khader, 11th c. BC) 𐤀, (Amurru, 11th c. BC) 𐤀, (Aḥīrām,ca. 1000 BC) 𐤀, 𐤀, (Jeḥīmilk, ca. 950 BC) 𐤀, (Elībaal, ca. 900 BC) 𐤀, (Tekke, ca. 900 BC) 𐤀, *Hismaitic* 𐫀, 𐫀, 𐫀, *Safaitic* 𐫀, 𐫀, 𐫀, 𐫀 <’> /ʔ/; *SHR* 𐫀 <ë> /ä, e, ē/, 𐫀, 𐫀 <ö> /ö, ő/; *TR* 𐫀 <A> /a, ä/

SFG-5: *Phoenician* (*Punic,* Motya, mid-6th c. BC) 𐤁, 𐤁 (Röllig 1995, 210–211) <’> /ʔ/ *SW* 𐤁, 𐤁, 𐤁, (Espanca) 𐤁 <o> /o/; *SE-Iberian* 𐤁, 𐤁, 𐤁 (Rodríguez Ramos 2004, 99) <o> /o/. *(i)* Rodríguez Ramos proposed that the P.-Hispanic 𐤁 <o> originated from the Phoenician <’> (2002, 192).

SFG-6: *Messapic* 𐌀, 𐌀, *Oscan* 𐌀, *Gallo-Greek* 𐌀 <a>

SFG-7: *Greek* (not later than 5th c. BC) 𐤀, *Lycian* 𐤀, (TL 5) 𐤀, (TL 33) 𐤀, 𐤀 (Kalinka 1901 *apud* Adiego 2015, 20–21), *S. Picene* 𐤀, *SE-Iberian* 𐤀, *NE-Iberian* 𐤀, 𐤀, *Celtiberian* 𐤀, *SHR* 𐤀 <a>. *(i)* The similarity of the shapes of the AH 𐤀, 𐤀, 𐤀 *19 <á> and the 𐤀, 𐤀 <a> has not been clarified.

SFG-8: *NE-Iberian, Celtiberian* 𐤀 <a>; *SHR* 𐤀 <a>

SFG-9: *Carian* 𐌀, *SE-Iberian* 𐤀, *Elymian* 𐤀, *Latin (epigraphic cursive)* 𐌀, (*cursive majuscule,* Pompeii) 𐤀, *Raetic* 𐤀, 𐤀, *Lepontic* 𐌀, 𐤀, *Gallo-Etruscan* 𐤀, 𐤀, *Camunic* 𐌀, *Runic (older fuþark)* 𐤀; 𐤀 <a>. *(i)* The Runic 𐤀 could be an autapomorphy.

SFG-10: *Elymian* 𐌀, *Raetic* 𐌀, 𐌀 <a>

SFG-11: *Parthian* 𐤀 <’> /a, ā/; *Sogdian* 𐤀, 𐤀 <’> /a, ā, ə/; *Syriac* 𐤀 <’>

SFG-12: *AH* 𐤀, 𐤀, 𐤀 *19 <á>; *Sidetic* 𐤀, 𐤀, 𐤀, 𐤀, 𐤀 <a> /a/; *TR* 𐤀, 𐤀, 𐤀, (manuscripts) 𐤀 <A> /a, ä/. *(i)* The possible relationship between SFG-11 and SFG-12 is unclear. The TR 𐤀 <A> could belong to SFG-11 and not SFG-12. It is noteworthy that there are some interesting, but maybe unrelated, orthographical features as follows. (a) The TR 𐤀, 𐤀 <A> also used as word separator. (b) According to Younger, the Cretan Hieroglyphic 𐤀 (Younger 2003-2012) is a phrase termination; however, Karnava (1999) handles this as a syllabogram and not a stiktogram. (c) The AH 𐤀 *450 <a> was also used as a word ending mark (Payne 2010, 81); however, it differs from the AH *19 <á> (SFG-12).

SFG-13: *P.-Canaanite* 𐤁 <b>; *Phoenician* (Byblos, 11th–10th c. BC) 𐤁 <b>; *P.-Hebrew* (late 8th c. BC) 𐤁 <b>; *Old Aramaic* (Zinjîrlû, late 9th–8th c. BC) 𐤁, (8th c. BC) 𐤁, (8th–7th c. BC) 𐤁 <b>; *SW* 𐤁; 𐤁, 𐤁, 𐤁 <p[e]> /p/; *NE-Iberian* 𐤁, 𐤁 <bi>; *Celtiberian* 𐤁, 𐤁, 𐤁, 𐤁, 𐤁 <bi>; *Greek* 𐤁, 𐤁, 𐤁, (Naxos, 8th–7th c. BC) 𐤁, (Argos) 𐤁 <β>; *Lycian* 𐤁 <b>; *Parthian* 𐤁, 𐤁 <b> /b, u̯/; *Sogdian* 𐤁, 𐤁 <b> /b, β/; *Hatran* 𐤁 <b>; *Syriac* 𐤁 <b, b̲>; *SR* 𐤁, 𐤁 <b[1]> /b, β/ (Vékony 2004, 315); *TR* 𐤁, 𐤁, 𐤁, 𐤁, 𐤁, 𐤁, 𐤁, 𐤁 <b[1]> /b/. *(i)* Glyph evolution: CT-11. *(ii)* For the possible evolution of the Greek 𐤁 <b> cf SFG-119.

SFG-14: *P.-Sinaitic* 𐤂, *P.-Canaanite* (Izbet Sartah, ca. 1100 BC) 𐤂, *Phoenician* 𐤂, 𐤂, 𐤂, 𐤂 <g>; *Old Aramaic* 𐤂, 𐤂, *Greek* 𐤂, 𐤂, 𐤂, 𐤂, 𐤂, *Elymian* 𐤂 <γ>; *P.-Umbrian* (Tolfa, ca. 530–525 BC) 𐤂 <c> (Urbanová 2003, 33; Bakkum 2009, 380); *Phrygian* 𐤂, 𐤂 <g>; *Etruscan* 𐤂 <c[e, i]> /k/; *Messapic* 𐤂, *Oscan* 𐤂, *Gallo-Greek* 𐤂, *I. Aramaic* 𐤂, 𐤂, *Parthian* 𐤂, *Middle Persian* 𐤂, *Hatran* 𐤂, *E. Cyrillic* 𐤂 <g>

SFG-15: *Greek* 𐤊, 𐤊, *Elymian* 𐤊 <γ>; *Lydian* 𐤊, 𐤊 <g>; *Etruscan* 𐤊 <c[e, i]> /k/; *Faliscan* 𐤊, 𐤊 <c>; *S. Picene* 𐤊 <c/g>; *Oscan* 𐤊, 𐤊, 𐤊, 𐤊, 𐤊, *Camunic* 𐤊, 𐤊 <g>; *Umbrian* 𐤊 <c/k> [k]; *Umbrian*

(late) 𐌶 <g>; *Runic (older fuþark)* ᚲ, ᛉ <k>; *Latin (archaic)* 𐌒 <c/g>; *Latin (classical)* G <g>; *NE-Iberian* 𐌺, 𐌶 <ge/ke>

SFG-16: *I. Aramaic* 𐡊, *Hebrew* 𐤊, 𐤣, *Parthian* 𐭣, *Palmyrene* 𐡊, *Nabataean* 𐢘 <g>; *Sogdian* (earlier than 4th c. AD) 𐼒 (Sims-Williams and Grenet 2007), (Ancient Letters) 𐼒 (Skjærvø 1996, 519), *(Manichean)* 𐫒, *(Christian)* 𐫒 <g, γ>; *Syriac* ܓ, ܟ, *Arabic* ڮ <ǧ>

SFG-17: *Greek* (*early minuscular,* 9th c. AD) 𐅃 <γ> (Taylor 1883, 154); *Galgolitic* 𐌶, 𐌶 <g>

SFG-18: *Lycian* 𐊗, 𐊘, (TL 5) 𐊘 (Kalinka 1901 *apud* Adiego 2015, 21) <g> /ɣ/; *TR* 𐊗, 𐊗, 𐊘, (manuscript) 𐊗, 𐊗 <ṅ> /ŋ/

SFG-19: *Madhabic* (Dadan) 𐤇, 𐤉, *Sabaic, Hasaitic, Dispersed ONA, Taymanitic, Dadanitic* 𐤇, *Dumaitic* 𐤄, *Taymanitic, Th. B* 𐤇, *Safaitic* 𐤇, 𐤄 (Macdonald 2015, 37), *Hismaitic* 𐤆 (Macdonald 2005, 82), 𐤇, 𐤃, *Th. C, Th. D* 𐤃, *Geʿez abjad* 𐍀, 𐍄 <d> /d/

SFG-20: *P.-Canaanite* (Izbet Sartah, ca. 1100 BC) 𐤃 (Cross 1989, 82) <d>; *Phoenician* 𐤃, 𐤃 (Sprengling 1931, 55), (Byblos, 11th–10th c. BC) 𐤃 <d>; *P.-Hebrew* (late 8th c. BC) 𐤃 <d>; *Old Aramaic* (10th–9th c. BC) 𐡃, 𐡃, 𐡃, 𐡃, (8th c. BC) 𐡃; (Deir 'Allā, around 800 BC) 𐡃; (8th–7th c. BC) 𐡃 <d>; *Greek* Δ <δ>; *Phrygian* Δ (Adiego 2007e, 3) <d>; *Lycian* 𐊃 (Adiego 2007e, 8) <d> /d/; *Faliscan* 𐌃, *Elymian* (5th c. BC) Δ, *S. Picene* 𐌃; *Oscan* Δ, *Messapic* 𐌃, Δ, Δ <δ, d> [d]; *Gallo-Greek* Δ, Δ <d> /d, t/; *SW* Δ <tᵘ> /t/; *SE-Iberian* Δ <tu> /tu/; *NE-Iberian* Δ <du/tu> /du, tu/; *Celtiberian* (Botorrita, Spain) 𐊃 (Eska 2008, 166–167), Δ, 𐊁 <tu>

SFG-21: *Greek* Δ <δ>; *Etruscan* (Marsiliana d'Albegna, 8th c. BC) 𐌃, 𐌃, (Veias, Caere, 7th c. BC) Δ <d> /t/; *Latin, Faliscan, Elymian, Messapic, Umbrian, Oscan* Δ <d, δ> [d]; *(i)* The Runic (older fuþark) 𐌃, (Jutland, ca. AD 160–350) 𐌃, 𐌃 (Looijenga 1997, 82–83) <Þ> may belongs to SFG-21.

SFG-22: *Greek (medieval cursive)* 𐆊 (Faulmann 1880, 171) <δ>; *E. Cyrillic* Д *dobro* <d>

SFG-23: *Greek (medieval cursive)* 𐆊 <δ>; *Glagolitic* 𐆊 <d>

SFG-24: *Lin. A* 𐘀 LA 01 <da>; *Lin. B* 𐘀 <da> /da/; *CM* 𐘀, 𐘀 CM 04 <ta?> (Valério 2016, 428); *CGk* (Common) 𐘀, (Paphian) 𐘀 (Olivier 2008, 617–618), 𐘀 <ta> /da, ta/; *Lydian* 𐤮, 𐤲 <d>. *(i)* Cf SFG-176. *(ii)* Cf Sidetic 𐤕 <t>.

SFG-25: *Lin. A* 𐘁 (Valério 2013, 15–17) LA 05 <to?>; *Lin. B* 𐘁 <to> /to, tʰo/; *CM* 𐘁 CM 13 / 𐘁 CM 78 <to?> (Valério 2016, 111–112, 430); *CGk* (ICS 172, *early*) 𐘁 (Valério 2016, 237), *(Common)* 𐘁; (Davis 2012, 38–61) *(Common)* 𐘁, *(Paphian)* 𐘁; (Olivier 2008, 617–618) (*Paphian,* 6th c. BC) 𐘁, 𐘁, 𐘁 (Valério 2016, 228) <to> /do, to/; *CGk* 𐘁 (Valério 2016, 230), *(Common)* 𐘁 (Davis 2012, 38–61) 𐘁, (*Paphian, late*) 𐘁 (Olivier 2008, 617–618) (*Paphian,* 6th c. BC) 𐘁, 𐘁 (Valério 2016, 228) <tu> /du, tu/. *(i)* The SFG-25 and SFG-26 are likely relatives. *(ii)* The SFG-25 and SFG-30 may be relatives.

SFG-26: *NE-Iberian* 𐊓, 𐊓, 𐊓 <do/to>; *Celtiberian* (Botorrita, Spain) 𐊓, 𐊓 <to>. *(i)* Cf SFG-25.

SFG-27: *SW* 𐌀, 𐌀 <tᵒ/tᵒ/ᵘ?> /t/; *SE-Iberian* 𐊀<tu> /tu/; *NE-Iberian* 𐊀, 𐊀, 𐊀, 𐊀 <du/tu>; *Celtiberian* 𐊀, 𐊀 <tu>. *(i)* The SFG-27, SFG-28, and SFG-29 are likely relatives.

SFG-28: *SE-Iberian* 𐊐 <tu> /tu/; *Celtiberian* 𐊐, 𐊐 <tu> /tu/. *(i)* Cf SFG-27 and SFG-29.

SFG-29: *Carian* 𐊫, 𐊫 <δ> /mḍ/d/ⁿt/; *SW* 𐊫 <tᵒ>; *Libyco-Berber* 𐤘; 𐤘 <T_3>; *TR* 𐊫, 𐊫, 𐊫, 𐊫, 𐊫, 𐊫, 𐊫, 𐊫, 𐊫, 𐊫, 𐊫 <t^1>. *(i)* The sound value /ⁿt/ of the Carian <δ> is supported by Kloekhorst (2008, 138–139). *(ii)* For the development of the TR glyphs, see CT-2 and CT-1. *(iii)* Cf SFG-27 and SFG-28.

SFG-30: *Greek* (Crete) Λ <δ>; *Phrygian* Λ <d>; *Libyco-Berber* 𐤋, 𐤋, 𐤋, 𐤋 <D>; *Sidetic* 𐤃 <d>;

Carian 𐊫, 𐊬, 𐊭, 𐊮 <d> /d/; *Celtiberian* 𐌍 <tu>; *CBR* 𐌃 <d> /d/ (Table 15 in Appendix); *SR* 𐌃, 𐌄 <d> /d, ð, ǰ/ (Vékony 2004, 243, 251, 267, 287, 294); *SR* 𐌅 <d> /d, ǰ/ (Vékony 2004, 253, 264, 287, 294). *(i)* Glyph evolution: CT-1. *(ii)* Cf SFG-25.

SFG-31: *P.-Sinaitic* (Serabit al Hadim, early 15th c. BC) 𐤘, 𐤙, 𐤚, 𐤛; (Sinai 358) 𐤟 (Colless 2010, 96) <h>; *AH* 𐤟 (Hawkins 1986, 370–371) *451 <hur>; *Messapic* 𐌚, 𐌛, 𐌜 <h?>; *SR* 𐌇, 𐌈 <h> /h/ (Vékony 2004, 287, 294); *CBR* 𐌈, 𐌇, 𐌊 <χ> /χ/ (Vékony 2004, 151)

SFG-32: *Madhabic* (Dadan) 𐪉, *Sabaic* 𐩨, 𐩩, *Hasaitic* 𐩩, *Dumaitic* 𐩵, *Taymanitic* 𐩵, 𐩩, *Dadanitic* 𐪈, 𐪊, 𐪋, *Th. D* 𐩩, 𐪁, *Th. C* 𐪁, *Th. B* 𐪈, 𐩩, *Hismaitic* 𐪉 (Macdonald 2005, 82), 𐪈, 𐪉 (King 1992, Figure 1 between pages 5 and 6), 𐩩, *Safaitic* 𐪋, 𐪌, 𐪍, 𐪉; 𐪉, 𐪌, 𐪋, *Ge'ez abjad* 𐌗 <h> /h/; *SR* 𐌄, 𐌉, 𐌊, 𐌋, 𐌍 <A> /a, ā, ä/ (Vékony 2004, 314); *CBR* 𐌄, 𐌗, 𐌉, 𐌉 <A> /à, a, ä, e/ (Table 15 in Appendix; Vékony 2004, 164, 185). *(i)* For the sound values see the comment in SFG-68. Moreover, in the Old Aramaic, the word-end <-'> and <-h> represented /-ā/. In the 10th–9th c., in the Old Aramaic the <-h> denoted /-ā/-t and /-ē/ (Segert 1978, 112–113). In the P.-Hebrew, the <h> denoted the word-ending /o/, /a/ or /e/ (Healey 1990, 35).

SFG-33: *Phoenician* (Inscription of King Kilamuwa, Zincirli, ca. 825 BC) 𐤄 <h>; *NE-Iberian* 𐌄, 𐌉 <e> /e/; *SR* 𐌄, 𐌅 <e> /ä, e/ (Vékony 2004, 314); *SHR* 𐌗, 𐌚 <e> /ä, ē/. *(i)* Cf SFG-38. *(ii)* The NE-Iberian 𐌄, 𐌉 <e> could be a direct variant of the NE-Iberian 𐌚 <e> (SFG-39) and not a direct descendant of the Phoenician 𐤄 <h>.

SFG-34: *P.-Canaanite* (Izbet Sartah, ca. 1100 BC) 𐤄, *Phoenician* (Byblos, 11th–10th c. BC) 𐤄, 𐤇, *P.-Hebrew* (end of 8th c. BC) 𐤄 <h>; *Phrygian* 𐌄, 𐌇, 𐌄, *Greek* (Athens, 8th–7th c. BC) 𐌇, (Corinth) 𐌇, 𐌄, *Etruscan* 𐌇, 𐌇, *Faliscan* 𐌄, *S. Picene* 𐌄, *Lemnian* 𐌇, *Messapic* 𐌄, 𐌄, *Venetic* 𐌇, 𐌉, *Camunic* 𐌇, *Elymian* 𐌄, 𐌄, *Raetic* 𐌇, 𐌋, 𐌋, *Lepontic* 𐌇, 𐌊, *Gallo-Etruscan* 𐌉, 𐌊, *Oscan* 𐌇, *Umbrian* 𐌉, 𐌊, 𐌇, *Latin* (*archaic)* 𐌄 <e, ε> /e/

SFG-35: *Greek* (8th–7th c. BC) 𐌉, 𐌉, 𐌇 <ε> /e/, *Lycian* 𐌄 <i> /i/; *Lydian* 𐌉, *P.-Campanian* (Sorrento) 𐌉, *S. Picene* 𐌄, 𐌉, *Oscan* 𐌉, 𐌆, 𐌄, *Elymian, Faliscan, Gallo-Greek, Umbrian, Latin (classical)* 𐌄, *Messapic* 𐌄, 𐌄 <e, ε> /e/

SFG-36: *Phrygian* 𐌋, 𐌋 (Young 1969, 262–268) <e> /e/; *Lydian* 𐌇, 𐌋, 𐌄 (Adiego 2007e, 7) <e> /e/ [e:]; *Camunic* 𐌇 (Morandi 2004, 476) <e>; *NE-Iberian* 𐌋, 𐌋, 𐌋 <e> /e/; *SR* 𐌅 <e> /ä, e/ (Vékony 2004, 314). *(i)* Presumably, the SW 𐌋 <h/H?> is also relative of the graphemes in SFG-36.

SFG-37: *Greek* (before 280 BC) 𐌄, 𐌔; (*minuscular,* 10th–11th c.) 𐌔; *(medieval cursive)* 𐌢; *Oscan* 𐌔 <ε, e>; *Messapic* 𐌔, *Umbrian* 𐌑, *Gallo-Greek* 𐌔, 𐌀, *Glagolitic* 𐌴 <e>, *E. Cyrillic* 𐌾 <e>

SFG-38: *Greek* (*cursive*, AD 701–718) 𐌀, 𐌁 <ε>; *Glagolitic* (Codex Zographensis, 10th–11th c.) 𐌶, 𐌴 <e>. *(i)* The similarity between the Glagolitic 𐌶, 𐌴 <e> and the SHR 𐌚 <e> (SFG-33) is maybe a homoplasy.

SFG-39: *Phoenician* (Sarepta, ca. 725 BC) 𐤅 <h>; *Greek* 𐌄 <ε>; *Lydian* 𐌄, 𐌅, 𐌄 <e>; *NE-Iberian* 𐌚, 𐌄, 𐌅, 𐌅 <e> /e/; *Celtiberian* 𐌄 <e>; *SHR* 𐌄 <e> /ä, ē/

SFG-40: *P.-Umbrian* (Tolfa, ca. 530–525 BC) 𐌌 <e> (Urbanová 2003, 33; Bakkum 2009, 380); *P.-Campanian* (Nuceria, second half of the 6th c. BC) 𐌌 <e> (MNAMON)

SFG-41: *Phoenician* (ca. 900 BC) 𐤉, 𐤄, *Old Aramaic* (Zinjîrlû, late 9th–8th c. BC) 𐤄, (8th c. BC) 𐤄, (Deir 'Allā, ca. 800 BC) 𐤄, (8th c. BC) 𐤄, 𐤄 <w>; *SW* 𐌇, 𐌇, (Espanca) 𐌇, *SE-Iberian* 𐌇 <u> /u/. *(i)* Cf SFG-44. *(ii)* Cf SFG-45.

SFG-42: *Greek* 𐌅, 𐌄 <ϝ> /u̯/ [w]; *Phrygian* 𐌅, 𐌄 <v> /w/; *Lycian* 𐌅 <w> /w/; *Lydian* 𐌀 <v> /v/; *Lemnian* 𐌄, 𐌆 <v>; *Etruscan* 𐌄 <v> /β, u̯/ [β]; *Raetic* 𐌀, 𐌀 <v> /v/; *Messapic* 𐌅, 𐌅, *Venetic* 𐌄,

Lepontic 𐌅, ꓷ <v>; *Umbrian* ꟻ, *Oscan* ꓷ <v> [w]; *Umbrian, Oscan* F, *Latin* ꟻ, F <f> [f]; *Runic (older fuþark)* ᛉ, ᚠ, ᚠ <f> /f/; *CBR* 𐌅 (Table 15 in Appendix; Hosszú and Zelliger 2014a, 186), *SR* 𐌅 <β> /β, v/ (Vékony 2004, 314)

SFG-43: *Greek* (8^{th}–7^{th} c. BC) Ꮍ, Ꮍ, Y, Ꮍ, Y, Y, Ꮍ, Ꮍ, Y; Ꮍ <υ> /u, ū/; *Greek (classical)* Y <υ> /u, ü/; *Phrygian* ᚴ, *Sidetic* Y, *Lydian* ᛚ, *Carian* Y, *Etruscan* (7^{th} c. BC) ᚴ, *Elymian* Y, *Messapic* Y, *Oscan* Y, *Gallo-Greek* Y, Y, ᚴ, ᚴ, ᛚ, Y, Y <u, υ> /u/; *Latin* (*archaic,* 4^{th}–2^{nd} c. BC) Y <v>

SFG-44: *I. Aramaic* (7^{th} c. BC) ᚿ, ᚿ, ᚿ, ᚿ, ᚿ; (6^{th}–4^{th}. c. BC) ᚿ, ᚿ, ᒣ, ᒣ, (Aśoka, around 250 c. BC) ᒣ, *Parthian* (Nisa, 1^{st} c. BC) ᚿ <w> /u̯, ō, ū/; *Hebrew* (1^{st} c. BC) ᚿ <w>; *Hatran* ᚿ <w>; *TR* ᐳ, ᚿ, ᚿ, ᐳ, *CBR* ᚿ (Vékony 2004, 151; Hosszú and Zelliger 2014a, 188), *SHR* Ͻ <W> /o, u/. *(i)* The possible ancestor of the Rovash ᐳ glyph is attested in the Aramaic script in 7^{th}–3^{rd} c. BC. The more characteristic Rovash ᚿ, ᚿ glyphs are attested from a narrower period, the 7^{th} c. BC. Consequently, Rovash most probably borrowed the Aramaic <w> in that time. The Rovash Ͻ was probably derived from ᐳ–like glyph by turning the short bars to obtain a closer shape (CT-2). *(ii)* Cf SFG-45.

SFG-45: *TR* ᚴ, ᛘ, ᚼ, *SHR* ᛋ, ⋜ <Ẅ> /ö, ü/. *(i)* Sebestyén (1915, 158) argued that the TR <Ẅ> is a descendant of the Greek (classical) Y <υ> /u, ü/ (SFG-43); *(ii)* In some Semitic scripts (e.g., Uyghur and Sogdian), the /ö, ü/ are represented by the ligature of the <y> and <w> (Erdal 2004, 42). Sims-Williams (1981, 359; 1989, 181; 1996, 313–314) demonstrated the Sogdian tradition of representing front rounded vowels *(ö, ü)* by the combination of <w> and <y>. Supposing the influence of the Sogdian script, the Rovash <Ẅ> could have been constructed of the Rovash ᚽ, ᛚ <i, y> (SFG-81) and ᐳ, ᚿ <W> (SFG-44) as follows: ᚴ < ᛚ + ᐳ; ᛘ < ᚽ + ᚿ; ᚼ < ᚽ + ᐳ; however, there is not direct evidence for this ligature-based evolution of the Rovash <Ẅ>. *(iii)* Erdal (2016) discovered use of the graphemes <o> for /ö/ and <u> for /ü/ as demonstrated in a Turkic text written with Brāhmī script in the IOL Toch 81 inscription (Maue 2008). According to the present author, a possible ancestor of the TR ᚴ, ᛘ, ᚼ <Ẅ> could be the Greek Ꮍ, Ꮍ <υ> /u, ū/, which could be used for representing /ö, ü/ as happened in the Brāhmī script; the glyph variants of the TR ᚴ, ᛘ, ᚼ <Ẅ> can be easily derived from the Greek Ꮍ, Ꮍ <υ>. In this solution, either the ligature forming or the /ü/ sound value of borrowed Greek Y <υ> have not to be assumed; therefore, based on *lex parsimoniae*, this lineage is the most probable. It is noteworthy that the glyphs of the SW ᚺ, ᚺ <u> (SFG-41) and the TR ᛘ, ᚼ <Ẅ> are very similar to each other. *(iv)* The Rovash ᛋ and ⋜ <Ẅ> are presumably variants of ᛋ <Ẅ>.

SFG-46: *P.-Sinaitic* ⟜, ♀ <w>; *AH* ♀ *280 <wa/i_9>. *(i)* The existence of this SFG is very tentative.

SFG-47: *Sabaic, Dispersed ONA, Taymanitic, Th. B, Hasaitic* ♀, *Dumaitic, Dadanitic* ♀, ♀, *Hismaitic* ♀, ᑫ, ♀, *Safaitic* ♀, ᚴ <y> /y/ [ç]; *Carian* ᚠ <ý> [ɥ]; *TR* ʔ, ᑫ, 9 <y^1> /y/ [j]; *TR* ♀, 𝟫, ᑭ, ᑫ <y^2> /y/ [j]

SFG-48: *P.-Umbrian* (Tolfa, ca. 530–525 BC) ↑ <f>; *Faliscan* ↑ <f>. *(i)* They may be relatives of the CGk (Paphian, 6^{th} c. BC) ⇑, ⇑ (Valério 2016, 228), (Common) ⇑ (Valério 2016, 230) <wo>. *(ii)* Cf SFG-49.

SFG-49: *NE-Iberian* ↑, ⋀, ⋒, ⦀, *Celtiberian* ↑ <u>; *S. Picene* ⋀ <ú> [uː]. *(i)* Cf SFG-48.

SFG-50: *P.-Sinaitic* ⥱, ⥱ <z/d̠?>; *Phoenician* ⥱, ⥱, I <z> /d͡z/; *P.-Hebrew* ⥱ <z>; *Old Aramaic* I, Z, I <z>; *Greek* (Crete, 8^{th}-7^{th} c. BC) I <ζ> /ds, sd/, /zd/ or /dz/ [d͡ʒ/d͡z]; *Oscan* I <z> [z, t͡s, d͡z]; *Libyco-Berber* ᚺ, H, I <Z_1>; *Dadanitic* ∀, ∀, H, *Taymanitic, Th. D* H, *Th. C,*

Hismaitic Ɪ, ⊢⊣, *Th. B* ⊤, *Safaitic* ⊤, ⫰, *Geʿez abjad* H <z> /z/

SFG-51: *Greek* (Athens, 8th–7th c. BC) I <ζ> /ds, sd/, /zd/ or /dz/ [d͡ʒ/d͡z]; *Sidetic* \ <z>; *Libyco-Berber* —, I <Z_2>; *I. Aramaic* (Aśoka, around 250 c. BC) \ <z>; *Parthian* (Nisa, 1st c. BC) I <z> /z, ž/; *Hebrew* (Qumran, 1st c. BC) I <z>; *Hatran* (shortly before AD 240) I <z>; *Sogdian* (earlier than 4th c. AD) \ <z>; *Syriac* (1st c. AD) \ <z>

SFG-52: *Phoenician* (Limassol [Cyprus], ca. 750 BC) ‡ <z>; *Greek* ⫲, ǂ <ζ> /ds, sd/, /zd/ or /dz/ [d͡ʒ/d͡z]; *Etruscan* ⫲ <z> /ts/; *Faliscan* ⫲, *Raetic* ⫲, *Lepontic* ‡, ‡, *Venetic* ⫲ <z>; *Umbrian* ⫲, ⫲, ⫲ <z> [t͡s]; *SHR* ⫲, ǂ, ⫽, ⫽ <ď> /ɟ/. *(i)* After 10th c. AD, in the Hungarian language, there was a /d͡ʒʲ/ > /ɟ/ change. Presumably, the origin of the Rovash ⫲, ǂ, ⫽ <ď> is the Greek ǂ <ζ>, and its ancestor is the P.-Sinaitic ═ <z/ḏ?> (SFG-50); however, the similarity between the Rovash ⫽ <ď> and the P.-Sinaitic ═ <z/ḏ?> is surely a homoplasy. *(ii)* Glyph evolution: CT-5.

SFG-53: *P.-Canaanite* ⧗, ⧗ <z>; *Madhabic* (JSMin 24) ⧗, *Sabaic (early musnad)* ⧗, *(early zabūr)* ⧗, *(middle musnad)* ⧗, *(middle zabūr)* ⧗ <z> (Macdonald 2015, 37, 39); *TR* (Yar Khoto grafitti no. 21) ⧗ (Erdal 1993, 91, 104–105); ⧗, ⧗, ⧗, ⧗, (Mendur-Sokkon IV) ⧗ (Konkobaev et al. 2015, 41) <z> (Kairzhanov 2014, 18). *(i)* Presumably, a ⧗ > ⧗, ⧗ shape transformation happened in TR, similarly to the ⧗ > ⧗ shape transformation (CT-8) in the Sabaic script. *(ii)* Cf SFG-59.

SFG-54: *CM* ⧗, ⧗, ⧗, ⧗, ⧗, ⧗ CM 107, ⧗, ⧗ CM 108 <za_2?/zo_2?/zi?>; *CGk* ⧗, ⧗, ⧗, ⧗, ⧗ <za?> /ɟja ↔ ʥa/? (based on Valério 2016, 227); *Carian*)(, ⧗, ⧗, (coins) ⧗, ⧗ <z> /sd/. *(i)* Cf SFG-59.

SFG-55: *Madhabic, Sabaic* H; *Hasaitic* H; *Dadanitic* ⧗, ⧗, ⧗, ⧗ (Macdonald 2010, 13–14), *Dumaitic* ⊟, *Taymanitic* H, *Th. B* ⧗, ⧗, ⧗ <ḏ> /ḏ/ [ð]. *(i)* Cf SFG-59. *(ii)* Glyph evolution: CT-7.

SFG-56: *Th. C* ⧗, *Hismaitic* ⧗, ⧗, ⧗, ⧗, *Safaitic* ⧗, ⧗, ⧗, ⧗, ⧗, ⧗ <ḏ> /ð/. *(i)* Cf SFG-58.

SFG-57: *NE-Iberian* ⧗, ⧗, ⧗ <da/ta> /da, ta/, ⧗ <ta>; *Taymanitic* ⧗, ⧗, *Dadanitic* ⧗, ⧗, ⧗, ⧗, *Th. C* ⧗, *Hismaitic* ⧗, ⧗, H <ṯ> /θ/; *SHR* ⧗, ⧗ <ť> /ť/ (cf Hung. /t/ > /ť/); *SR* ⧗ (Vékony 2004, 253, 264, 315) <t> /t/. *(i)* Cf Lycian)(, ⧗, ⧗ <θ> /θ/ (Adiego 2007e, 8; Mechert 2008b, 49)

SFG-58: *P.-Campanian* (Nuceria) ⧗, (Sorrento) ⧗ <s>; *Camunic* ⧗, ⧗, ⧗ <z>; *Runic (older fuþark)* ᛉ <z> /z/. *(i)* The graphemes in SFG-56 and SFG-58 are maybe relatives. *(ii)* Glyph evolution: CT-5.

SFG-59: *SHR* ⧗, ⊟, ⧗, ⧗, ⧗, ⧗, ⧗, ⧗, ⧗ Z <z> /z/; *SHR* ⧗, ⧗, ⧗, □ <č>; *SR* ⧗, ⧗ <č> /č/; *SR* ⧗, ⧗ <č> /s/ (Vékony 2004, 314). *(i)* The closest relative of the Rovash ⊟ <z> could be one of the following: CM ⧗ 108 <za_2?/zo_2?/zi?> (SFG-54), S. Semitic ⊟ <ḏ> (SFG-55) or S. Semitic ⧗ <z> (SFG-53). *(ii)* Glyph evolution: CT-5.

SFG-60: *Lin. B* ⧗ <zo> [t͡s, d͡z] (Valério 2016, 216); *AH* ↑, ⧗, ⧗ *376 <zi> /tsi/; *Phrygian* ⧗, ↑, ⧗ <↑> /tʃ/ts/; *Lydian* ↑ <c> /z/dz/dʒ/ (Melchert 2004, 602–603; 2008b, 58) or /tˢ/ (Valério 2008, 130); *Carian* ↑ <τ> /tš/; *Sidetic* ↑ <ts>; *SHR* ↑ <c> /č, ts/. *(i)* Valério (2008, 130–131.) proposed (referring to Melchert 2004) the relation between the AH ↑ <zi> and Phrygian ↑ /tʃ?/. According to Adiego (2004, 302–303) the Lydian ↑ <c> and the Carian ↑ <τ> are relatives.

SFG-61: *Madhabic* Ψ, *Sabaic* Ψ, Ψ, *Hasaitic* Ψ, *Dispersed ONA, Taymanitic* Y, *Dadanitic, Th. C* ⧗, *Hismaitic* ⧗, ⧗; ⧗, ⧗, *Th. B* Y, *Geʿez abjad* ⧗ <ḥ> /ħ/. *(i)* Cf SFG-62.

SFG-62: *Greek* (red) Y, ↓, Ѵ, Ψ <kh> /kʰ/; *Etruscan* Y, Ѵ <χ> [kʰ]; *Raetic* Y, V, ⅄ <χ> /ch/; *Lepontic* ↓, Ѵ <χ>; *Venetic* Y <χ> [g]; *Camunic* Y, Y <χ> [g]; *Gallo-Etruscan* Y, ⧗ <χ>. *(i)* Cf SFG-61.

SFG-63: *P.-Sinaitic* 𐤢, 𐤡, 𐤣, *P.-Canaanite* 𐤡, 𐤈, 𐤢, *Phoenician* 𐤡, 𐤄, 𐤅, 𐤇, *P.-Hebrew* 𐤇, *Old Aramaic* (Zinjîrlû, late 9th–8th c. BC) 𐡄, 𐡇, (8th c. BC) 𐡀 <ḥ>; *SW* 𐤇, 𐤄 <h>; *Greek* 𐤇 (Woodard 2014, 37), 𐤄, 𐤋, 𐤌, 𐤍, 𐤏 <η> /ē, h/; *Elymian* (5th c. BC) 𐤍, 𐤄 <H>; *Etruscan* 𐤄, 𐤄, *Raetic* 𐤌, 𐤍, 𐤌, 𐤌, *Faliscan* 𐤃, 𐤄, *Venetic, Oscan, Latin (archaic)* 𐤄, *Messapic* 𐤄, *Umbrian* 𐤄, 𐤌 <h>; *Runic (Anglo-Saxon)* 𐤇, *(older fuþark)* (Oostum, The Netherlands) 𐤇 (Looijenga 1997, 73) <h>. *(i)* Cf SFG-65.

SFG-64: *Greek* (Naxos, 8th–7th c. BC) 𐤏 <η>, *Carian* (Memphis, Sinuri, Stratonikeia) 𐤏 <e>

SFG-65: *Greek* (6th c. BC) 𐤇, 𐤇 <ε> /ē, h/; *Lemnian* (6th c. BC) 𐤉 <h>; *Elymian* (5th c. BC) 𐤇, 𐤍 <H>; *Carian* (Mylasa) 𐤇 <e>, *Messapic* 𐤇 <ē, h>, *Oscan* 𐤇 <η, ē>, *Gallo-Greek* 𐤇, 𐤈, 𐤍 <h>; *Runic (older fuþark)* 𐤇 <h>; *Latin, Oscan, Umbrian* 𐤇 <h>. *(i)* Cf SFG-63.

SFG-66: *NE-Iberian* 𐤇, 𐤇, 𐤋, 𐤋, 𐤄 <o>; *Celtiberian* (Botorrita, Spain) 𐤇 (Eska 2008, 166–167) <o>. *(i)* The SFG-66 and SFG-63 could be related if vowel value of SFG-66, since in the early age, the grapheme <h> was occasionally used to denote /o/ in the Old Aramaic and P.-Hebrew scripts (Healey 1990, 35). *(ii)* Glyph evolution: CT-6.

SFG-67: *Greek (medieval cursive)* 𐤊, 𐤗 <kh>; *Glagolitic* 𐤗 <x>

SFG-68: *P.-Sinaitic* 𐤨, 𐤨 <ḫ> /χ/; *SHR* 𐤧, 𐤧, 𐤩 <h> /h/; *SHR* 𐤧, 𐤩 <e> /e/; *TR* 𐤩 (Tekin 2003, 23); 𐤧, 𐤤, 𐤧 (Kairzhanov 2014, 17) <e> /e/; *SR* 𐤩 <e> /e/ (Vékony 2004, 287, 294). *(i)* Using the <h> or <ḫ> for representing /ē/ was specific for the Greek (similarly in Lydian, Lycian, Phrygian); however, in the Old Aramaic, the <-h> was also used for /-ē/ (Segert 1978, 113) in the 10th–11th c. BC. Therefore, the value /e/ of the Rovash grapheme 𐤧, 𐤩 <e> could originate from the Greek, Lydian, Lycian, Phrygian, or the Old Aramaic, but not from the later Aramaic. *(ii)* Glyph evolution: CT-1, CT-5. *(iii)* Cf SFG-32.

SFG-69: *P.-Canaanite* (Izbet Sartah, ca. 1100 BC) 𐤛 <ṭ>; *Phoenician* 𐤘, 𐤕, 𐤕 <ṭ/ṯ> /tˢ/; *Old Aramaic* 𐤗, 𐤕, 𐤛 <ṭ/ṯ>; *Greek* 𐤊, 𐤔, 𐤈, 𐤗, (cursive) 𐤈, 𐤚, *Etruscan* 𐤊, 𐤘, *Lemnian* 𐤗, *Messapic* 𐤀, 𐤔, 𐤈, *Venetic* 𐤁, 𐤊 <θ>; *Sidetic* 𐤊, 𐤏 <th> /θ/; *Oscan* 𐤔 <f> [f]; *Gallo-Greek* 𐤔, 𐤔, 𐤉, 𐤉 <θ> /t͡s/; *Safaitic, Th. D, Hismaitic* 𐤙 <ḍ> [ðˢ]; *SW* 𐤀 <tⁱ> /d, t/; *SE-Iberian* 𐤀 <ti>, *NE-Iberian* 𐤀, 𐤃, 𐤔, 𐤈, 𐤗, *Celtiberian* 𐤗 <de/te> /de, te/; *TR* 𐤗 <dª/ᵉd>; 𐤏, 𐤊, 𐤎, 𐤎, 𐤊, 𐤊 <nd/nt>; *SHR* 𐤃, 𐤉, 𐤁, 𐤃, 𐤈, 𐤁 <f>. *(i)* In the AH and Aegean syllabaries, the /n/ before consonant was not written (Fischer 2001, 75). Maybe that is why the TR <nd/nt> represented /n/ plus consonant (cf SFG-87 and SFG-100). *(ii)* The Greek glyphs 𐤗, 𐤈 were typical even in the 7th c. BC (McCarter 1975); however, in the 4th c. BC these glyphs did not appear in the surviving Greek inscriptions (Thompson 1912, 144–145). Consequently, these glyphs were borrowed by Rovash before the 4th c. BC. See comment *(ii)* in SFG-166.

SFG-70: *Madhabic* 𐤄, 𐤄, *Sabaic (early musnad)* 𐤄, *(early zabūr)* 𐤄, *(middle zabūr)* 𐤄, *Hasaitic* 𐤄, *Dumaitic* 𐤇, *Th. C* 𐤋, *Safaitic* 𐤇, 𐤇 (Macdonald 2015, 30, 37), *Taymanitic* 𐤄, 𐤄, *Dadanitic* 𐤌, *Th. B* 𐤇, 𐤋, 𐤋, 𐤋 <ḍ> /ḍ/ [ðˢ]. *(i)* Glyph evolution: CT-5. *(ii)* Cf SFG-71.

SFG-71: *SW* 𐤇, 𐤇, 𐤋, (Espanca) 𐤋 <tᵉ> /d, t/; *SW* 𐤋, 𐤌 <tª> /t/. *(i)* The graphemes in SFG-70 and SFG-71 could be indirect relatives.

SFG-72: *Greek* (*cursive,* 2nd c. BC – 9th c. AD) 𐤈, (*minuscular,* 9th c. AD) 𐤈, (*late uncial,* 9th c. AD) 𐤈, (*minuscular,* 10th–11th c. AD) 𐤈 <θ>; *Glagolitic* 𐤈, 𐤈 <f>; *E. Cyrillic* 𐤈, 𐤈 <f>. *(i)* See comment *(ii)* in SFG-166.

SFG-73: *Greek* 𐤈 <θ>; *Lemnian* 𐤌 <θ>; *SE-Iberian* 𐤀 <ti>, *NE-Iberian* 𐤀, 𐤀, 𐤀, 𐤀, 𐤀, *Celtiberian* (Botorrita, Spain) 𐤀, *(Eastern)* 𐤀 <de/te>

SFG-74: *Parthian* 𐫢, 𐫟 <ṭ/t̲> /t^{s}/; *TR* 𐭕, 𐭖, 𐭗, 𐭘, 𐭕 <d^{1}> /d/. *(i)* In the Parthian and Middle Iranian languages, in intervocalic position a voicing occurred: /p/ > /b/, /t/ > /d/ and /k/ > /g/ (Skjærvø 1996, 519). This could be a reason why the Parthian <ṭ/t̲> represented /d/ in the Rovash.

SFG-75: *Lin. A* 𐘂 (Valério 2013, 15–17), 𐘃 LA 57 <ja>; *Lin. B* 𐘃, 𐘄 <ja>; *CM* 𐘃, 𐘅, 𐘆 CM 69, 𐘇, 𐘈, 𐘉 CM 71 <ja?>; *CGk* 𐘊, 𐘋, 𐘌, 𐘍 <ja>; *TR* 𐭉, 𐭊, 𐭋, 𐭌, 𐭍, 𐭎 <y^{1}> /y/ [j]; *SR* 𐭉, 𐭋 <y> /y/ [j] (Vékony 2004, 315); *CBR* 𐭋 <y> /j, i̯/ (Table 15 in Appendix; Vékony 2004, 164); *SHR* 𐭉, 𐭊, 𐭊, 𐭊, 𐭊, 𐭊 <ĺ> /j/. *(i)* Cf Carian 𐊮, 𐊯, 𐊰, 𐊱, 𐊯, 𐊲, 𐊱, 𐊳, 𐊴, 𐊵, 𐊶, 𐊷, 𐊸 (Adiego 2007a, 209–210, 508) <i> and Lydian 𐤁 <y> /i/. The Lydian <y> /i/ is an unstressed allophone of [i] (Melchert 2008b, 59).

SFG-76: *P.-Umbrian* (Tolfa, ca. 530–525 BC) 𐌉 <í> (Urbanová 2003, 33; Bakkum 2009, 380); *S. Picene* 𐎇, 𐎈, 𐎉, 𐎊, 𐎋 <e> (MNAMON)

SFG-77: *Phoenician* 𐤄, *Aramaic* 𐡄, 𐡄, 𐡄, *Old Aramaic* (8th c. BC) 𐡄, *I. Aramaic* (7th c. BC) 𐡄, (6th c. BC) 𐡄, 𐡄 <y> /y/ [j]; *Lydian (archaic)* 𐤄 <i> (Woudhuizen 1984–1985a, 93). *(i)* Cf SFG-81.

SFG-78: *Lydian (archaic)* 𐤉 <i> /i/ (Woudhuizen 1984–1985a, 93); *SW* 𐤉, (Espanca) 𐤉, *SE-Iberian* 𐤉, 𐤉 <i> /i/; *NE-Iberian* 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, *Celtiberian* (Botorrita, Spain) 𐤉, 𐤉 <i>

SFG-79: *Sidetic* 𐤉 (Adiego 2007e, 14), 𐤉; 𐤉, 𐤉 (Woudhuizen 1984–1985b, 117) <i> /i/; *NE-Iberian* 𐤉, 𐤉 <i> /i/

SFG-80: *Greek* 𐤉, 𐤉, 𐤉, 𐤉 <ι> /i, ī/; *Lydian* 𐤉 (Woudhuizen 1984–1985a, 93) <i> /i/; *Phrygian* 𐤉, 𐤉 <y> /j/; *Sidetic* 𐤉 <j>; *Libyco-Berber* 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, 𐤉, 𐤉 <Y/I>; *Runic (older fuþark, Anglo-Saxon)* 𐤉 <ï> /ij/

SFG-81: *I. Aramaic* (7th c. BC) 𐡉, 𐡉, (6th c. BC) 𐡉, 𐡉, 𐡉, 𐡉, (5th/4th c. BC) 𐡉, *Hebrew* (Qumran, 1st c. BC) 𐡉 <y>; *TR* 𐭉, 𐭉, 𐭉 <i, y> /i, ï/; *SHR* 𐭉, 𐭉 <j> /i, j/; *CBR* 𐭉 <i> /i/ (Vékony 2004, 164); *SR* 𐭉, 𐭉 <i> /e, i, ï/ (Vékony 2004, 314). *(i)* Cf SFG-45. *(ii)* Cf SFG-77.

SFG-82: *Hatran* 𐣠 <y>; *Sogdian* 𐼹 <y> /y, ē, ï/, *Nabataean* 𐢚 <y>. *(i)* The similarity of the SFG-80 and SFG-82 is probably homoplasy due to the lack of known historical and geographical relationship.

SFG-83: *Palmyrene* (Palmyra, 2nd c. AD) 𐡉 <y>; *Middle Persian (Inscriptional)* 𐭩, *(Psalter)* 𐭩, *(Early Cursive Pahlavi)* 𐭩, *(Book Pahlavi)* 𐭩, 𐭩, 𐭩 <y> /y, ě, ǐ, ǰ/; *TR* 𐭩 <y^{2}> /y/; *CBR* 𐭩 (Table 15 in Appendix), *SR* 𐭩, 𐭩, 𐭩 <i> /i, ï/ (Vékony 2004, 314).

SFG-84: *Greek (medieval uncial)* 𐅁, 𐅁 (Faulmann 1880, 171), (*late uncial,* 9th c. AD) 𐅁 (Taylor 1883, 154) <ι>; *Glagolitic* 𐰆, 𐰇 *iže* <i>; *E. Cyrillic* 𐅁 *iže* <i>

SFG-85: *Glagolitic* 𐰋 *jerь* <ь>; *E. Cyrillic* 𐅁 *jerǐ* <ь>. *(i)* Cf *Glagolitic* 𐰆 <i> (SFG-84).

SFG-86: *Glagolitic* 𐰝 *ju* <j>; *E. Cyrillic* 𐅁 *ju* <j>. *(i)* Cf *Greek* (*cursive,* AD 701–718) 𐅁, 𐅁, 𐅁 (Thompson 1912) <ι>.

SFG-87: *Lin. A* 𐘀, 𐘀, 𐘀 LA 77 <ka?>; *Lin. B* 𐘀 <ka> /ka, ga, k^{h}a/; *CM* (ENKO Atab 001, not later than 1525–1425 BC) 𐘀 CM0 09 (Valério 2016, 186) <ka?>; *S. Picene* 𐎀 <q>; *NE-Iberian* 𐎀, 𐎀, 𐎀 <gu/ku> /gu, ku/, 𐎀, 𐎀 <gu>, 𐎀, 𐎀 <ku>; *Celtiberian* 𐎀 <ku>; *Carian* 𐊫, 𐊫, 𐊬 <q> /q/; *Th. B* 𐊬, *Th. C* 𐊬, *Safaitic* 𐊬 <g>; *Runic (older fuþark)* 𐊬 <ŋ> /ŋ/; *TR* 𐊬, 𐊬, 𐊬 <ṅ> /ŋ/. *(i)* Cf Th. D 𐊬 <g?>. *(ii)* In the AH and Aegean syllabaries, the /n/ before consonant was not written (Fischer 2010, 75). Maybe that is why the Runic <ŋ> and the Rovash <ṅ> could represent a nasal sound. Cf SFG-69 and SFG-100.

SFG-88: *CM* 𐘁, 𐘁, 𐘁, 𐘁, 𐘁, 𐘁 CM 25 <ka?>; *CGk* 𐘁 <ka> /ga, ka, k^{h}a/; *TR* 𐭒, 𐭒 <k^{5}/^{w}k^{w}> /q/.

(i) The possible relation of the Greek (red) 𐤧, 𐤭 <kh> to SFG-88 has not been clarified.

SFG-89: *SE-Iberian* 𐬀, 𐬁 <ga>; *NE-Iberian* 𐬂, 𐬃, 𐬄, 𐬅, 𐬆, 𐬇, 𐬈, 𐬉, 𐬊 <ga/ka>; *Celtiberian* (Botorrita, Spain) 𐬋, 𐬌, 𐬍, 𐬎, 𐬏, 𐬐 <ka>; *SHR* 𐬑, 𐬒, 𐬓 <g>

SFG-90: *AH* (KARKAMIŠ A14b, 10th c. BC) 𔓐 (Hawkins 2000, 83–86), 𔖵, 𔖶 *315 <kar>; *CBR* 𐬔 <k> (Vékony 2004, 165); *TR* 𐬔, 𐬕, 𐬖, 𐬗, 𐬘, 𐬙 <k^4, $^{\text{w}}$k$^{\text{w}}$>. *(i)* It is noteworthy that a relationship between the Rovash 𐬔 <k> and Rovash 𐬗 <q> (SFG-101) could exist if supposing the CT-9 characteristic transformation. However, the altering sound values of the Rovash 𐬔 <k> and 𐬗 <q> implies different origins. Cf CT-5.

SFG-91: *Lin. A* 𐘣, 𐘤, 𐘥, 𐘦, 𐘧, 𐘨 LA 74 <ze?> /ce, ɟe ↔ tse, dze/; *Lin. B* 𐘪, 𐘫 <ze>; *CM* 𐚸, 𐚹, 𐚺, 𐚻, 𐚼 CM 112 <k/ze?> /ce, ɟe ↔ tse, dze/?; *CGk* 𐚽, 𐚾, 𐚿, 𐛀, 𐛁 <ke> [k$^{\text{j}}$e, g$^{\text{j}}$e, k$^{\text{hj}}$e ↔ ce, ɟe, c$^{\text{h}}$e] (Valério 2016, 256). *(i)* The SFG-91 is based on Valério (2016, 253–256).

SFG-92: *Lin. A* 𐘣, 𐘤, 𐘥, 𐘦, 𐘧, 𐘨 LA 74 <ze?> /ce, ɟe ↔ tse, dze/; *CM* 𐚸, 𐚹, 𐚺, 𐚻, 𐚼 CM 112 <k/ze?> /ce, ɟe ↔ tse, dze/ (Valério 2016, 256); *CBR* 𐬜, 𐬝, 𐬞, 𐬟, 𐬠 <z> /z/ (Table 15 in Appendix; Vékony 2004, 164). *(i)* Both the similarity between the Lin. A 𐘤 and the Rovash 𐬝 glyphs and the sound relation between the Lin. A <ze?> and Rovash <z> are tentative proposals and need more evidence. *(ii)* Glyph evolution: CT-5.

SFG-93: *Phoenician* (Nora, ca. 900 BC) 𐤊, (Kilamuwa, ca. 825 BC) 𐤊, (Limassol, ca. 750 BC) 𐤊; *Old Aramaic* (10th–5th/4th c. BC) 𐡊 (Gibson 1975), (8th c. BC) 𐡊; (Deir 'Allā, around 800 BC) 𐡊 <k> /k, χ/; *SW* (Espanca) 𐬡, 𐬢, 𐬣 <k$^{\text{e}}$> /k/; *NE-Iberian* 𐬡, 𐬢 <ke>; *Greek* (8th–7th c. BC) 𐌊, 𐌊, 𐌊, 𐌊, 𐌊, 𐌊 (Healey 1990, 37), *Phrygian* 𐌊, *Lemnian* 𐌊, *Lydian* 𐤊, *Elymian* 𐌊, 𐌊, 𐌊, *Raetic* 𐌊, 𐌊, 𐌊, 𐌊, *S. Picene* 𐌊, 𐌊, *Faliscan* 𐌊, *Lepontic* 𐌊, 𐌊, 𐌊, *Venetic* 𐌊, 𐌊, 𐌊, 𐌊, *Camunic* 𐌊, *Messapic* 𐌊, *Oscan* 𐌊, 𐌊, 𐌊, *Latin (archaic)* 𐌊 <k, κ> /k/; *Lycian* 𐊊 <k> /k</ (Melchert 2008a, 48) or /c/ (Adiego 2007e, 8); *Etruscan* 𐌊 <k$^{\text{a}}$> /k/; *Umbrian* 𐌊, 𐌊, 𐌊, 𐌊, 𐌊, *Gallo-Etruscan* 𐌊, 𐌊, 𐌊, 𐌊, 𐌊, *Gallo-Greek* 𐌊, 𐌊 <k> [g, k]

SFG-94: *Phoenician* (Amurru, 11th c. BC) 𐤏, 𐤏, (Aḥīrām, Byblos, ca. 1000 BC) 𐤏, 𐤏, (Jeḥīmilk, Byblos, ca. 950 BC) 𐤏, (Elībaal, Byblos, ca. 900 BC) 𐤏 <k>; *NE-Iberian* 𐬤, 𐬥, *Celtiberian* 𐬦 <ge/ke>; *SR* 𐬧 (Vékony 2004, 315), *TR* 𐬨, 𐬩, 𐬪, 𐬫 <k^2> /k/

SFG-95: *NE-Iberian* 𐬬, 𐬭 <ge/ke>; *TR* 𐬬 <k^2> /k/. *(i)* Glyph evolution: CT-1. *(ii)* The similarity to the Runic (Anglo-Saxon) 𐬭 <η> /ŋ/ is presumably a homoplasy.

SFG-96: *NE-Iberian* 𐬮 <ge/ke>; *TR* 𐬮, 𐬯, (Mendur-Sokkon IV) 𐬰 (Konkobaev et al. 2015, 41) <g^2>

SFG-97: *Lin. A* 𐘭, 𐘮, 𐘯, 𐘰 LA 67 <ki?>; *Lin. B* 𐘯; 𐘰 <ki> /gi, ki, k$^{\text{h}}$i/; *CM* 𐚮, 𐚯, 𐚰, 𐚱, 𐚲, 𐚳 CM 70 <ki?> (Valério 2016, 436, 442); *CGk* 𐛂, 𐛃, 𐛄, 𐛅, 𐛆, 𐛇, 𐛈, 𐛉 <ki> /gi, ki, k$^{\text{h}}$i/; *Carian* 𐊧, 𐊨, 𐊩, 𐊪, 𐊫, 𐊬 <k> /k/; *CBR* 𐬴 <k> /k/ (Hosszú 2013, 38–39); *SHR* 𐬴, 𐬵, 𐬶 <k> /k/; *TR* 𐬷, 𐬸, 𐬹, (manuscript) 𐬺 <k^3/$^{\text{i}}$k$^{\text{i}}$> /q/

SFG-98: *Greek* (Crete, Sikinos, Phrygia) 𐌚 (Jeffery 1961, 35–37, 39–40) <kh?>; *Sidetic* 𐌘 <g>; *SE-Iberian* 𐬻 <ki>; *NE-Iberian* 𐬻, 𐬼, 𐬽, 𐬾, 𐬿, 𐭀, 𐭁, 𐭂, 𐭃, 𐭄, 𐭅, 𐭆, 𐭇 <gi/ki> /gi, ki/; *Celtiberian* (Botorrita, Spain) 𐬽, 𐬻 <ki/ci>; *SHR* 𐬿, 𐭀 <k> /k/; *CBR* 𐬿 <k> /k/ (Vékony 2004, 164); *SHR* 𐭇 <γ> /γ/; *TR* 𐭁, 𐭂, 𐭃, 𐭄, 𐭅, (manuscript) 𐭆 <k^1> /q/; *TR* 𐭈 <g^1> /γ/; *CBR* 𐬽, 𐭁, 𐭂, 𐭄 <γ> /γ/ (Vékony 2004, 164, 192; Hosszú and Zelliger 2014a, 186); *SR* 𐭀, 𐭆, 𐭇, 𐭉, 𐭊, 𐭋, 𐭌, 𐭍 <g> /g, γ/ (Vékony 2004, 315). *(i)* The Lycian 𐊵, 𐊷 <β/K> /k$^{\text{w}}$/? (Adiego 2007e, 8; Melchert 2008a, 48) is maybe also member of SFG-98.

SFG-99: *Lin. A* 𐘬 LA 70 <ko?>; *Lin. B* 𐘬, 𐘭 <ko> /ko, go, k$^{\text{h}}$o/; *CM* 𐚴 CM 21 <ko?>; *CGk* 𐛊, 𐛋, 𐛌, 𐛍, 𐛎, 𐛏 <ko> /go, ko, k$^{\text{h}}$o/; *Libyco-Berber* 𐤼, 𐤽, 𐤾, 𐤿, 𐥀, 𐥁, 𐥂, 𐥃, 𐥄 <G>; *SW* 𐬊 <k$^{\text{a}}$> /k/;

SE-Iberian <ka>; *NE-Iberian* <ka>; *Celtiberian* <ka/ca>

SFG-100: *CM* , CM 15 <ko?> (Valério 2016, 430, 442); *AH* (Payne 2010, 14), (SÜDBURG) , , *423 <ku> /gu, ku/; *Lycian* , , <k>; *NE-Iberian* , , , , , <gu/ku> /gu, ku/; *Celtiberian* (Botorrita, Spain) , <ku/cu>; *S. Picene* , , , <q>; *Safaitic* , , , (Macdonald 2015, 37) <g>; *Runic (older fuþark)* , <ŋ> /ŋ/; *SHR* , , <k> /k/; *SR* , <k[1]> /q/ (Vékony 2004, 315); *TR* , , <ṅ> /ŋ/. *(i)* Cf SFG-87. Presumably, both CM CM0 09 <ka?> (SFG-87) and *CM* , CM 15 <ko?> (SFG-100) are ancestors of the P.-Hispanic, Ancient Italic, and Rovash graphemes in SFG-87 and SFG-100; which is an example of glyph-level reticulation.

SFG-101: *CGk* (Paphian) , <ko> /go, ko, k[h]o/ *Carian* , ?, (E.Me 30) <ɣ> /g/, (coins) , (M33) , <ɣ?> (Adiego 2007a, 483–509); *SW* , <k[o]> /k/, *SE-Iberian* <go/ko> /go, ko/, *NE-Iberian* , , <go/ko> /go, ko/, , , <ko> /ko/; *Celtiberian* (Botorrita, Spain) , , <ko/co>; *Th. C* , *Hismaitic* <g>; *SHR* <χ> /χ/; *CBR* <q> /q/ (Vékony 2004, 165; Hosszú and Zelliger 2014a, 186), *SR* <k[1]> /q/; *SR* <k[5], [w]k[w]> /q/ (Vékony 1992a, 542; 2004, 315). *(i)* The Celtiberian <ko> could be autapomorphy; however, it is more probable that this open shape also existed in other cognate scripts, cf Aegean <ko>. *(ii)* The Rovash <q> and S. Semitic <g> are probably not homoplasies. *(iii)* Cf *Th. D* <g?>. *(iv)* Cf CT-8 and CT-9.

SFG-102: *Dadanitic* , , , *Dispersed ONA* , *Taymanitic* , , <g>; *SR* , <g> /g, ɣ/ (Vékony 1992a, 542). *(i)* These may be the one-loop version of the or shapes in SFG-101; since the probable relative S. Semitic <g> surely has a relationship with the shapes in SFG-101.

SFG-103: *Lin. A* , LA 44 <ke?>; *Lin. B* <ke> /ke, ge, k[h]e/; *CM* , , , , , , CM 110 <ke/u?>; *CGk* , , , , , , <ku> /gu, ku, k[h]u/; *Libyco-Berber* , <Q>; *Lycian* , , , <q>; *TR* , , , , , , *SR* , <g[1]/[1]ɣ> /ǧ, ɣ/ (Vékony 1992a, 542). *(i)* Glyph evolution: CT-3.

SFG-104: *Carian* , <k̂/χ (Simon 2008, 459–460)> /c?/k[j]?/k[w]?/; *Lydian* <q> /k[w]/? (Adiego 2007e, 7; Melchert 2008b, 57); *Runic (older fuþark)* <g> [g, ɡ, ɣ, j].

SFG-105: *AH* *176; , , (BABYLON 1) (Payne 2010, 121) *175 LINGUA <la>; *Carian* , <l>. *(i)* The relation of the AH *175 and the Carian <l> is uncertain, cf SFG-162.

SFG-106: *P.-Sinaitic* , , (Sprengling 1931, 55) <l>; *P.-Canaanite* (Izbet Sartah, ca. 1100 BC) (Rollston 2008, 84) <l>

SFG-107: *P.-Sinaitic* <l>; *Madhabic, Sabaic, Dispersed ONA, Dumaitic, Dadanitic, Th. C, Th. B., Hasaitic* , *Taymanitic, Hismaitic* , , *Th. D* , *Safaitic* , *Ge'ez abjad* <l>; *Greek* , , <λ>; *Phrygian* , *Lydian* , *Lemnian* , *S. Picene* , *Camunic* , *Messapic* , *Raetic* , *Venetic* , *Runic (older fuþark, Anglo-Saxon, younger/Danish fuþark)* , *SW* , (Espanca) , *SE-Iberian* , *NE-Iberian* , *Celtiberian* (Botorrita, Spain) , <l>

SFG-108: *Phoenician* , , , *P.-Hebrew* , *Old Aramaic* (ca. 800 BC) ; (7th c. BC) , , , *I. Aramaic* , *Middle Persian* , , *Syriac* , *Arabic* <l>; *Sogdian* (Ancient Letters) <δ> (Skjærvø 1996, 519); *Greek* , <λ>; *Faliscan* , *Camunic* , *Etruscan* , *Raetic* , , *Lepontic* , , *Venetic* , *Oscan* , *Umbrian* , , *Latin (archaic)* , *SR* <l> (Vékony 2004, 315); *TR* , , , , <l[1]>

SFG-109: *Greek* (Ionia, Corinth) , *Elymian* , , <λ>; *Oscan* <λ, l>; *Lycian* (TL 29) , (TL

5) 𐤋 (Kalinka 1901 *apud* Adiego 2015, 14, 21), *NE-Iberian* 𐌋, *Messapic* 𐌋, 𐌛, *Gallo-Greek* 𐌋, 𐌛, 𐌚, *SHR* 𐌋, *SR* 𐌋 <l> (Vékony 1992a, 542). *(i)* The possible relationship of the Lycian 𐤋 <l> and the 𐤉 CM 011, CGk 𐌚, and SHR 𐌋 <l> in SFG-162 is unclear.

SFG-110: *Greek* (cursive, AD 824–830) 𐌚 <λ>; *Glagolitic* 𐌛 <l>

SFG-111: *Greek* (cursive, AD 701–718) 𐌙, 𐌝, 𐌞 (Thompson 1912) <λ>; *E. Cyrillic* 𐌠 <l>

SFG-112: *P.-Sinaitic* 𐌊, 𐌊, 𐌉, *P.-Canaanite* (mid-11th c. BC) 𐌉, *Phoenician* 𐌉, (Byblos, 11th–10th c. BC) 𐌉, *Old Aramaic* (Tell Fekheriye, 9th c. BC) 𐌉, *Greek* (Crete, 8th–7th c. BC) 𐌑, *SW* 𐌉, 𐌉, 𐌉, 𐌉, *P.-Umbrian* (Tolfa, ca. 530–525 BC) 𐌉, 𐌉, *SR* 𐌉, 𐌉 (Vékony 2004, 315), *CBR* 𐌉 <m> (Table 15 in Appendix)

SFG-113: *Phoenician* (Kilamuva-stele, Sam'al, around 820 BC) 𐤌, 𐤌 (Lemaire and Sass 2013, 125), 𐤌, 𐤌, *Old Aramaic* (8th c. BC) 𐤌, (Deir 'Allā, around 800 BC) 𐤌, (10th–7th c. BC) 𐤌 (Gibson 1975), *Greek* 𐌑, (8th–7th c. BC) 𐌑, (8th–7th c. BC) 𐌑, (8th–7th c. BC) 𐌑 <μ>, *Phrygian* 𐌑, *Lydian* 𐌑, *Lemnian* 𐌑, *Etruscan* (7th c. BC) 𐌑, *Raetic* 𐌑, 𐌑, 𐌑, *Lepontic* 𐌑, *Camunic* 𐌑, *Faliscan* 𐌑, *Venetic* 𐌑, *Latin (archaic)* 𐌑, *Carian* 𐌍, 𐌅 <m>

SFG-114: *P.-Hebrew* (8th c. BC) 𐤍, *Old Aramaic* (8th–3rd c. BC) 𐤍, *Phoenician* (Karatepe, ca. 700 BC) 𐤍, (Byblos, 5th–4th c. BC) 𐤍, *I. Aramaic* (7th–5th/4th c. BC) 𐤍, 𐤍, *Parthian* 𐤍, 𐤍, 𐤍, 𐤍, 𐤍, *Hatran* 𐤍 <m>

SFG-115: *I. Aramaic* 𐡌, *Hebrew* ם, *Middle Persian (inscriptional)* 𐭬, 𐭬, 𐭬, 𐭬, *(Psalter)* 𐮍, 𐮍, *(Book Pahlavi)* 𐮍, 𐮍, *Syriac* ܡ, ܡ, ܡ, *Sogdian* 𐼥, 𐼥 <m>

SFG-116: *CGk* (Paphian, 6th c. BC) 𐌌, 𐌌, 𐌌, 𐌌 (Valério 2016, 228, 278) <me>; *Libyco-Berber* ⵣ, ⵣ, ⵔ, ⵔ, ⵓ <P/F> /f/p?/; *NE-Iberian* 𐌏, 𐌏 <ba>, 𐌁, 𐌅, 𐌁, 𐌏, 𐌏, 𐌏, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌, 𐌌 <be>, *Celtiberian* 𐌏, 𐌏, 𐌏, 𐌏, 𐌏, 𐌅 <be/pe>; *TR* 𐌏, 𐌏, 𐌏, 𐌏, 𐌏, 𐌏, 𐌏 <b^2> /b, u̯/; *CBR* 𐌏 <b> /u̯/ (Table 15 in Appendix); *TR* 𐌏, 𐌏, 𐌏, 𐌏, 𐌏, 𐌏, ◇, (Mendur-Sokkon IV) 𐌏 (Konkobaev et al. 2015, 41) <m>. *(i)* The TR 𐌏 <b^2> ~ 𐌏, 𐌏, 𐌏 <m> could be related to the Old Turkic onset [b] > /m/ change (Erdal 2004, 62, 74), cf SFG-118. The TR 𐌏 <m> probably originated from the TR 𐌏 <b^2> by turning the shape 𐌏 with -90°, cf CT-10. *(ii)* The similarity between the TR 𐌏 <m> and the Middle Persian 𐮍, 𐮍, 𐮍, 𐮍 <m> (SFG-115) is likely a homoplasy.

SFG-117: *TR* 𐌌 (Gabain 1941) 𐌌, 𐌌, 𐌌 (Kairzhanov 2014, 17) <m>; *SHR* 𐌌, 𐌌, 𐌌, (Constantinople, AD 1515) 𐌌 <mb>. *(i)* The relationship between TR <m> and SHR <mb> was proposed by Németh (1934) and Vékony (2004). *(ii)* Presumably, the glyphs in SFG-117 are ornamented versions of the glyphs in SFG-116, cf TR 𐌏 <m> (SFG-116). *(iii)* The glyphs in SFG-117 are maybe comparable to the *SW* (Espanca) 𐌌, 𐌌; 𐌌, (Stele of Mestras) 𐌌 (Correia 1996, 105) <p^{i}?> (Valério 2008, 125–126).

SFG-118: *CGk* 𐌌, 𐌌 <mi>; *Lydian* (650–600 BC) 𐌌 <m> (Woudhuizen 1984–1985a, 98); *Madhabic, Sabaic* 𐩣, *Dispersed ONA* 𐩣, 𐩣, *Dumaitic, Taymanitic* 𐩣, *Dadanitic* 𐩣, 𐩣, *Th. B* 𐩣, 𐩣, *Hasaitic* 𐩣, 𐩣, *Geʿez abjad* 𐌌, *SHR* 𐩣, 𐩣, 𐩣, 𐩣 <m>. *(i)* Cf AH ◇ (Hawkins 1986, 370–371), 𐌌 (Payne 2010, 14) *419 <mí>.

SFG-119: *Phrygian* 𐌁, 𐌁 <b>; *Greek* 𐌁, 𐌁, 𐌁, 𐌁 (Jeffery 1961, 23), (Crete, 8th–7th c. BC) 𐌁, (Argos) 𐌁 <β>; *Carian* 𐌁, 𐌁, 𐌁, (coins) 𐌁, 𐌁 <p> /p/; *Lycian* 𐌁 (Adiego 2007e, 8) <b>; *Lydian* 𐌁 <b> /p/; *Etruscan* 𐌁 <b> [p]; *P.-Campanian* (Nuceria) 𐌁 <b>; *Elymian* 𐌁, 𐌁 <β>; *Umbrian* 𐌁, 𐌁, *Oscan* 𐌁, 𐌁, 𐌁, 𐌁, *Latin* 𐌁, *Messapic* 𐌁, 𐌁, *Gallo-Greek* 𐌁 <b>; *Runic (older fuþark, Anglo-Saxon, younger/Danish fuþark)* 𐌁, *(older fuþark)* 𐌁, 𐌁, (Oostum, The Netherlands) 𐌁 (Looijenga 1997,

6, 73) <b>; *TR* 𐠂 (Kairzhanov 2014, 17) <b[1]>. *(i)* It is possible that the Greek 𐠂 <b> was not invented from the glyph 𐠃 (SFG-13), but it indirectly originates from the *CGk* 𐠷, 𐠹 <mi> (SFG-118). In this case, this is an example for glyph-level reticulation. *(ii)* The TR 𐠂 <b[1]> is related to the graphemes in SFG-119, or the SHR 𐠂 <m> (SFG-118) ~ TR 𐠂 <b[1]> correspondence originates from the Old Turkic onset [b] > /m/ change (Erdal 2004, 62, 74), cf SFG-116. *(iii)* Valério (2016, 282–284) pointed out that languages without phonemic /m/ typically possess /b/ that—depending on its position—can be pronounced allophonically as [b], [m], or prenasalised [^{m}b]; the realization of the sound depends on if the following vowel is plain or nasalized. Valério also mentioned the possibility that the language of CM possessed a sound that varied between [b] and [m]. This opinion supports that the SFG-118 and SFG-119 could be relatives. *(iv)* Glyph evolution: CT-5. *(v)* Cf SFG-120.

SFG-120: *NE-Iberian* 𐠳 <m>, 𐠅, 𐠇, 𐠕, 𐠖, 𐠅, 𐠖, 𐠎 <ḿ/m̃>; *Celtiberian* 𐠹, 𐠇, 𐠳, 𐠳 <ḿ>. *Sidetic* 𐡀, 𐡂 <m>, *Umbrian* 𐠛 <m>; *Libyco-Berber* ⊔, ∪, ∨ <M>. *(i)* Presumably, the graphemes in SFG-120 are related to the CGk 𐠷 <mi> in SFG-118.

SFG-121: *Madhabic, Sabaic, Dispersed ONA, Dumaitic, Taymanitic, Dadanitic, Th. B, Th. C, Th. D* 𐠠, *Safaitic* 𐪉, 𐪊, 𐪋, 𐪌, 𐪍, 𐪎, 𐪏 (Macdonald 2015, 31, 33, 37), *Hismaitic* 𐪎, 𐪍, 𐪉, *Geʿez abjad* 𐠠 <b>

SFG-122: *CM* 𐠑, 𐠒, 𐠓, 𐠔 CM 73 <mo?>; *CGk* 𐠫; (Paphian, 6th c. BC) 𐠫, 𐠪, 𐠬, 𐠭, 𐠮, 𐠯, (Paphian, late) 𐠭, 𐠫, 𐠰, 𐠱 (Valério 2016, 278) <mo>; *SW* 𐠊 <p°>; *SE-Iberian* 𐠯 <bo?>; *NE-Iberian* 𐠮, 𐠯, 𐠰 <bu> /bu/; *Celtiberian* (Botorrita, Spain) 𐠈, 𐠯 <bu/pu>; *Libyco-Berber* ⊙, ⊡, ⊕, ⊕, ⊞ <B>; *TR* 𐠈, 𐠉, 𐠈, ⊙, O, *SR* O (Vékony 2004, 243, 251, 315) <m>. *(i)* Note that the shapes of the 𐠔 CM 73 <mo?> and Libyco-Berber ⊞ <B> are close to each other. *(ii)* The *AH* 𐠰 *362 <má> maybe belongs to SFG-122.

SFG-123: *CM* 𐠖, 𐠗, 𐠘, 𐠙, 𐠚, 𐠛 CM 39/49 <mu?>; *CGk* 𐠵, 𐠴, 𐠶, 𐠷, 𐠸, 𐠹 <mu>; *SW* 𐠋 <bu>; *SE-Iberian* 𐠋, *NE-Iberian* 𐠌, 𐠌, 𐠌, 𐠌, 𐠌, 𐠌, 𐠌, 𐠌, *Celtiberian* (Botorrita, Spain) 𐠌 (Eska 2008, 166–167), 𐠌, 𐠌 <bo>; *SHR* X, *CBR* X (Vékony 2004, 164), *SR* X <b> (Vékony 1992a, 542)

SFG-124: *P.-Canaanite* 𐤍, *Phoenician* 𐤍, *P.-Hebrew* 𐤍, *Old Aramaic* (8th c. BC) 𐡍, *I. Aramaic* (5th/4th c. BC) 𐡍, *Parthian* 𐡍, *CBR* (Nagyszentmiklós) 𐲍, 𐲎 <n> (Table 15 in Appendix)

SFG-125: *Madhabic* (Dadan) 𐪈, 𐪉, *Sabaic* 𐩢; *Dispersed ONA* 𐠞, *Dumaitic* 𐠞, 𐠞, *Taymanitic* 𐠞, 𐠞, *Dadanitic* 𐠞, *Th. D* 𐠎, *Hasaitic* 𐪉, 𐪈, *Geʿez abjad* 𐠞 <n>. *(i)* The difference between SFG-125, SFG-126, SFG-127, and SFG-128 is very small, and it is difficult to distinguish them.

SFG-126: *Greek* 𐠍, *Phrygian* 𐠍, *Lydian* 𐠍 (Adiego 2007e, 7; Melchert 2008b, 57), 𐠍 (Swiggers and Jenniges 1996, 283), *Etruscan* 𐠍, *Faliscan* 𐠍, *Lemnian* 𐠍, *Lepontic* 𐠍, *Raetic* 𐠍, 𐠍, *Venetic* 𐠍, *Camunic* 𐠍, *Messapic* 𐠍, *Gallo-Etruscan* 𐠍, 𐠍, 𐠍, *Latin (archaic)* 𐠍, *Oscan* 𐠍, *Elymian* 𐠍, *SW* 𐠍, 𐠍, *SE-Iberian* 𐠍, *NE-Iberian* 𐠍, *Celtiberian* 𐠍 <n/ν>. *(i)* See comment *(i)* in SFG-125.

SFG-127: *Greek* 𐠎, *Dispersed ONA* 𐠎, *S. Picene* 𐠎, 𐠎, *Lycian* 𐠎, 𐠎, 𐠎, *Etruscan* 𐠎, *Messapic* 𐠎, *Elymian* 𐠎, 𐠎, 𐠎, *Raetic* 𐠎, *Faliscan* 𐠎, *Oscan* 𐠎, 𐠎, *Gallo-Greek* 𐠎, N, *Umbrian* 𐠎, 𐠎, *Lepontic* 𐠎, *NE-Iberian* 𐠎, *Celtiberian* 𐠎 <n/ν>. *(i)* See comment *(i)* in SFG-125.

SFG-128: *Dispersed ONA* 𐠏, *Etruscan* 𐠏, *Messapic* 𐠎, 𐠏 <n>; *Elymian* 𐠎, *Gallo-Greek* 𐠏 <n/ν>. *(i)* See comment *(i)* in SFG-125.

SFG-129: *Oscan* 𐠏, *Umbrian* 𐠏, *Greek* (*cursive,* 601–640) 𐠏 (Thompson 1912), *E. Cyrillic* 𐠏 <n/ν>

SFG-130: *Greek* (*early minuscular,* 9th c. AD) 𐠏, 𐠏 (Thompson 1912), (*cursive,* AD 701–718) 𐠏

<ν>; *Glagolitic* Ⱀ, ⱀ <n>

SFG-131: *Phoenician* 𐤍, /,), *Hebrew* נ, נ, *Parthian* 𐭍, ـــ, ـ, ﻧ, ا, ׀, *Sogdian* (Ancient Letters, early 4[th] c. AD) 𐼌, *Middle Iranian (Psalter)* 𐮋, *(Book Pahlavi)* ׀, *Hatran* 𐣭, *Palmyrene* 𐡮, *Nabataean* 𐢓, 𐢒, *Syriac* ܢ (individual), ܢ (ending), ܢ (beginning, middle), *Sogdian (Christian)* ܢ, ܢ <n>. *(i)* Cf SFG-135.

SFG-132: *AH* 𔐀, 𔐁, 𔐂 *35 <na>; *SR* ϛ, ϟ, ϡ (Vékony 2004, 251, 267, 294, 315) <n>. *(i)* Cf Sidetic 𐤍 <n>.

SFG-133: *Lin. A* 𐘂, 𐘂, 𐘂 LA 24 <ne?>; *Lin. B* 𐀚, 𐀚, 𐀚, 𐀚, 𐀚; 𐀚 <ne>; *CM* 𒾐, 𒾐, 𒾐, 𒾐 CM 02, 𒾐, 𒾐, 𒾐, 𒾐 CM 34, 𒾐, 𒾐 CM 56 <ne?>; *CGk* 𐠜, 𐠜, 𐠜, 𐠜, 𐠜, 𐠜, 𐠜 <ne>; *TR* 𐰤, 𐰤, 𐰤, 𐰤, 𐰤, 𐰤, 𐰤 <n[2]> /n/. *(i)* Glyph evolution: CT-3.

SFG-134: *CM* (Ugarit) ⏀ (Valério 2016, 106–108) CM 02 <ne?>; *Carian* (Kaunos, Stratonikeia) Φ <ñ>

SFG-135: *AH* ᑐ (Hawkins 2010, 184, 188–189), ᘐ (Payne 2010, 119), ᘐ (Yakubovich 2015a, 12), ⊂ (Anders 2012), ᘓ (Payne 2010, 14), ⋐ (Payne 2010, 116) *411 <ni>; *TR*), (<n[1]> /n/; *SHR*) <n>; *CBR*) <n> /n, ñ/ (Vékony 2004, 164); *SR*) (Vékony 1992a, 542) <n> /n/. *(i)* Probable homoplasies: glyphs in SFG-131, since their glyphs are similar to the glyphs in SFG-135; however, in the scripts in SFG-131, dextrograde writing is impossible, and dissimilarly, in SFG-135, the glyphs have two opposite versions, e.g., the Rovash), (<n, n[1]>. Another difference, that all glyphs in SFG-135 are arched while certain glyph variants in SFG-131 are straightened: /, |. From this it follows that the arch ()) is only part of the visual identity of SFG-135, and not of SFG-131.

SFG-136: *CM* 𐀛, 𐀛; 𐀛, Ѱ, 𐀛 CM 65/67/99/100 <ni?> (Valério 2016, 435–436, 442); *CGk* 𐠝, 𐠝, 𐠝 <ni>; *Carian* V, Y <n>

SFG-137: *AH* 𔖐, 𔖐 *395 <nú>; *Lycian* 𐊖, 𐊖 <ñ> /n̥/

SFG-138: *P.-Canaanite* 𐤎, 𐤎 <s>; *Phoenician* 𐤎, 𐤎, 𐤎, 𐤎, 𐤎, 𐤎, (Lachish letters, 6[th] c. BC) 𐤎; (Byblos, 5[th]–4[th] c. BC) 𐤎 <s> /t͡s/; *P.-Hebrew* 𐤎, 𐤎 <s>; *Old Aramaic* 𐡎, 𐡎, 𐡎, 𐡎 <s> /s/; *SW* ǂ <s>, *SE-Iberian* ǂ, ǂ <s>; *Libyco-Berber* ⵌ, H, ⵌ <S_1>; *Greek* Ξ, ǂ, Ξ <ξ> /ks/. *(i)* Other members of SFG-138 could be: Elymian Ξ <ξ?> and Lydian Ξ, Ξ <τ> /ts/? (Adiego 2007e, 7, Melchert 2008b, 57–58).

SFG-139: *I. Aramaic* Δ, ◻ (Faulmann 1880, 171), *Hebrew* ס, ס, *Nabataean* 𐢖, 𐢖 <s>

SFG-140: *Parthian* 𐭎, *Palmyrene* 𐡯, *Hatran* 𐣮, *Sogdian* (Ancient Letters) 𐼍 <s>. *(i)* The SFG-140 can be relative of the SFG-144 or the SFG-139.

SFG-141: *Lin. A* Y (Valério 2013, 15–17) LA 31 <sa?>; *Lin. B* 𐀭, 𐀭, 𐀭, 𐀭, 𐀭, 𐀭 <sa>; *CM* Y, (RASH Atab 004) Y (Valério 2013, 19–20) CM 82 <sa?>; *CGk (Paphian)* Y <sa>; *SR* Y, T <s> /s/ (Vékony 2004, 315); *TR* 𐰽, 𐰽 <s[1], š>; *TR* 𐰞 <l[1]>, 𐰞, (manuscript) 𐰞 (Gabain 1941) <l[2]>. *(i)* The TR 𐰽, 𐰽 <s[1], š> and 𐰞 <l[1]> are relatives according to Róna-Tas (10). *(ii)* The close relationship of the Old Turkic /s/ and /š/ was discussed by Erdal (2004, 102).

SFG-142: *Madhabic* ሰ; *Sabaic, Dispersed ONA, Dumaitic, Hasaitic, Taymanitic, Th. B, Th. C, Th. D* ሰ, *Dadanitic* ሰ, ሰ, ሰ, ሰ (Macdonald 2010, 13–14), *Hismaitic* ሰ, ሰ, *Safaitic* ሰ, ሰ, ሰ, ሰ, ሰ <s[1]> [ʃ] (Macdonald 2004, 496, 499); *Geʿez abjad* ሰ <s[1]/s>. *(i)* Glyph evolution: CT-2.

SFG-143: *Hismaitic* ⋋, λ, ʎ, ⤙, *Safaitic* ⤚, ∠, λ, V <s[1]> [ʃ] (Macdonald 2004, 496, 499); *TR* 𐰲, 𐰲 <č> /č, ǰ/. *(i)* Cf the Old Turkic /š/ ~ /č/ (Erdal 2004, 103). *(ii)* SFG-142 and SFG-143 are certainly glyph variants of each other.

SFG-144: *CGk* <sa>; *AH* , *415 <sa>; *Safaitic* , , , (Macdonald 2005, 82) <s[1]> [ʃ]; *SHR* <š>; *TR* , <š[1], š[2]>, <š>; *SR* <š> (Vékony 2004, 315)

SFG-145: *Lin. A* , LA 09 <se?>; *Lin. B* <se>; *CM* CM 44 <se?> (Valério 2016, 433, 442); *CGk* , , , , , , , , <se> /se/ (Valério 2016, 206); *TR* , , , , , (J) , , , , , <s[1], š> /s, š/; *SR* <š> (Vékony 2004, 315); *TR* < s[1], š[1], s[2], š[2], r[2]>; *SR* <š> (Vékony 2004, 315); *SHR* <ž> /š, ʒ/; *TR* , <[i]č[i], [i]ğ> /č, ǰ/. *(i)* Erdal (2004, 102) discussed the close relationship of the Old Turkic /s/ and /š/. *(ii)* Erdal (2004, 103) described several cases of the /š/ ~ /č/ alternation.

SFG-146: *Greek* (*early minuscular,* 9th c. AD) <ω> /ō/; *Greek* (*cursive*, 3rd c. AD) , (*cursive,* AD 302–359) <o> /o, ō/; *Glagolitic* *onъ* <o>; *SHR* (Vargyas, 12th–13th c. AD) <o> /o/

SFG-147: *E. Cyrillic* (9th–10th c. AD) *onsъ (жсъ)* <ǫ>; *CBR* (Nagyszentmiklós) <ǫ> /ǫ/ (Table 15 in Appendix)

SFG-148: *Lin. A* LA 03 <pa?>; *Lin. B* <pa> /ba, pa, p[h]a/; *CM* CM 06 <pa?>; *CGk* , , <pa> /ba, pa/; *SHR* , <p> /p/. *(i)* Cf SFG-149.

SFG-149: *SW* <p[a]> /p/; *SE-Iberian* <be?>; *SHR* , <p> /p/. *(i)* The glyphs in SFG-149 are probably relatives of *SHR* , <p> in SFG-148; however, a relationship to SFG-153 or SFG-154 is also possible.

SFG-150: *Lin. A* , LA 56 <pa$_2$?>; *Lin. B* <pa$_3$>; *CM* CM 72b <pa$_2$?>; *SW* <p[u]> /p/

SFG-151: *Phoenician* , *Old Aramaic* , , , , *I. Aramaic* (7th c. BC) , , , , (6th c. BC) , (4th–3rd c. BC) , *Greek* , *Lycian* , , *Etruscan* , , *Umbrian* , , *Faliscan* , , *Raetic* , <p>; *Lepontic* <P> /b, p/; *Gallo-Etruscan* , <p> /b, p/; *SHR* , *TR* , *SR* <p> /p/ (Vékony 2004, 315). *(i)* Cf SFG-152.

SFG-152: *Carian* , , <b> /b/; *SE-Iberian* , *NE-Iberian* , , <ba>; *Celtiberian* , <ba/pa>. *(i)* Cf SFG-151.

SFG-153: *Lin. A* , , , LA 50 <pu?>; *Lin. B* , <pu> /bu, pu, p[h]u/; *CM* , , , CM 41 <pu?>. *(i)* Cf SFG-149.

SFG-154: *Lin. A* LA 29 <pu$_2$?>; *Lin. B* , , <pu$_2$>; *CM* (?), , , , , CM 37 <pu?/so?> (Valério 2016, 432, 442); *CGk* , , , , , <pu> /bu, pu/. *(i)* Cf SFG-149.

SFG-155: *NE-Iberian* <ś>; *Etruscan* <ś> [ʃ]; *Raetic* <ś> /ś/; *Lepontic* , <ś>; *Camunic* , <ś>; *Gallo-Etruscan* (4th–2nd c. BC) , <š>; *Libyco-Berber* , <S$_2$/S$_1$/S>

SFG-156: *P.-Canaanite* (Izbet Sartah, ca. 1100 BC) , *Phoenician* (Byblos, 11th–10th c. BC) , *P.-Hebrew* , *Old Aramaic* (10th–5th/4th c. BC) (cf SFG-158), *SW* , , *SE-Iberian* , *NE-Iberian* , , *Celtiberian* , *Greek* (8th–7th c. BC) , , *Phrygian* , , *Lemnian* , *Lycian* , *Lydian* , *Faliscan* , *Etruscan* , *Messapic* , , *Venetic* , *S. Picene* , , *Elymian* , , *Raetic* , *Gallo-Etruscan* , *Latin (archaic)* , , *Oscan* , *Gallo-Greek* , , , , <r>; *Umbrian* <ḍ/ř>, *Glagolitic* , *E. Cyrillic* <r>

SFG-157: *Greek* (8th–7th c. BC) <ρ>; *Etruscan* , *P.-Campanian* (Nuceria) , *Raetic* , , , *Lepontic* , , , *Venetic* (6th–1st c. BC) , , *Camunic* , *Oscan* (Etruscan-like, 4th–1st c. BC) , , *Umbrian* (Etruscan-like, 4th–1st c. BC) , , *Gallo-Etruscan* (4th–2nd c. BC) , *SW* (Espanca) , , *NE-Iberian* , , , <r>

SFG-158: *Old Aramaic* (middle 7th c. BC) (cf SFG-156), *I. Aramaic* (middle 8th c. BC) , (7–5th/4th c. BC) , , , (Aśoka, ca. 250 BC) , *Hatran* , , *Sogdian* <r>; *TR* , , , , , , , , <r[1]>; *SR* , (Vékony 2004, 315), *SHR* , , , , , <r>. *(i)* The Rovash glyph is attested

to in Aramaic from 8th–5th/4th c. BC; therefore, borrowing into the presumably common ancestor of the Rovash scripts called Proto-Rovash could happen in this period.

SFG-159: *Lin. A* 𐘀, 𐘁, 𐘂 LA 60 <ra?>; *Lin. B* 𐘂 <ra> /la, ra/; *CM* 𐘐 CM 87 <la?> (Valério 2016, 438, 442); *CGk* 𐘐, 𐘑, 𐘒, 𐘓 <la>; *Sidetic* 𐘔 (Adiego 2007e, 14) <l>. *(i)* Glyph evolution: CT-10.

SFG-160: *CM* 𐘕, 𐘖, 𐘗 CM 75 <ra>; *CGk* 𐘘, 𐘙, 𐘚, 𐘛, 𐘜 <ra>; *NE-Iberian* 𐘝, 𐘞, 𐘟, 𐘠, 𐘡, 𐘢, 𐘣, 𐘤 <ŕ>; *Celtiberian* (Botorrita, Spain) 𐘣, 𐘝, 𐘞, 𐘥, 𐘦, 𐘠, 𐘧, 𐘨, 𐘩 <r/ŕ>; *Libyco-Berber* 𐘝, 𐘪 <R>; *CBR* 𐘫, 𐘬 (Table 15 in Appendix; Hosszú and Zelliger 2014a, 188), *SR* 𐘭 <r> /r/ (Vékony 2004, 154, 314). *(i)* NE-Iberian <ŕ> is a liquid (Valério 2008, 130) or a trill (Ferrer i Jané 2013, 448). *(ii)* Glyph evolution: CT-1.

SFG-161: *AH* 𐘮, 𐘯, 𐘰, 𐘱 *383 <ra/i>; *SHR* 𐘲, 𐘳 <r>. *(i)* The similarity to the S. Semitic 𐘴 <r> (SFG-164) is presumably a homoplasy.

SFG-162: *Lin. A* 𐘵 LA 27 <re?>; *Lin. B* 𐘵 <re> /le, re/; *CM* 𐘶 CM 011, 𐘷, 𐘸, 𐘹, 𐘺, 𐘻, 𐘼 CM 24 <le?>; *CGk* 𐘽, 𐘾 <le>; *SHR* 𐘿, 𐙀, (Constantinople, AD 1515) 𐙁, 𐙂, 𐙃, 𐙄 <l>. *(i)* The SHR 𐙄, 𐙁 <l> and the Carian 𐙅, 𐙆 <l> (SFG-105) are comparable. *(ii)* Cf SFG-105. *(iii)* Cf SFG-109. (*iv*) The relationship between the glyphs 𐘵 and 𐘶 are discussed by Valério (2016, 266).

SFG-163: *CM* 𐘇 CM 33 <re?>; *CGk* 𐙇, 𐙈, 𐙉, 𐙊, 𐙋, 𐙌 <re> /re/; *Carian* (Kildara, Sinuri, Stratonikeia) 𐙍, (Memphis, Kaunos) 𐙎, (Memphis, E.Me 14) 𐙎, (Memphis, E.Me 37) 𐙏, (bronze lion, ca. 500 BC, E.xx 7, sinistrograde) 𐙏, (Tralleis) 𐙐 <r>; *Sidetic* 𐙑, 𐙐 <r>. *(i)* The Sidetic 𐙑, 𐙐 <r> may belong to SFG-164.

SFG-164: *Madhabic* (Dadan) 𐙒, 𐙓, *Sabaic* 𐙔, 𐙕, 𐙖 (Sprengling 1931, 55), *Hasaitic* 𐙔, *Dispersed ONA* 𐙔, 𐙗, 𐙘, *Dumaitic* 𐙔, *Taymanitic* 𐙔, 𐙘, ↄ, *Dadanitic* 𐙔, 𐙘, *Hismaitic* 𐙔, ↄ, ⊂ (Macdonald 2005, 82), *Th. B* 𐙔, 𐙗, *Safaitic* 𐙔, 𐙙; 𐙚, 𐙛, 𐙜, 𐙝, 𐙞, 𐙟, 𐙠 (Macdonald 2015, 37), 𐙡 *Th. D* 𐙢, *Th. C* 𐙣, 𐙢 <r>; *SR* 𐙤 <r2> /r/ (Vékony 2004, 315). *(i)* Cf SFG-161. *(ii)* Cf Carian (Memphis) 𐙥 <ŕ> /ri?/.

SFG-165: *Lin. A* 𐘊, 𐘋, 𐘌 LA 26 <ru?>; *Lin. B* 𐘌 <ru> /lu, ru/; *CM* 𐘋, CM 010, 𐘊 CM 28 <lu?>; *Lydian* (550–500 c. BC) 𐤬, 𐤭, 𐤮 <λ>; *TR* 𐤮 <r1> /r/, 𐤭, 𐤯 (manuscript) 𐤰, 𐤱, 𐤲 <r2> /r/

SFG-166: *Madhabic* 𐙦, *Sabaic (early musnad)* 𐙦, *(early zabūr)* 𐙧, *(middle musnad)* 𐙨 (Macdonald 2015, 39), *Th. B, Hasaitic* 𐙦, *Safaitic* 𐙩, 𐙪, 𐙫, 𐙬, 𐙭, 𐙮 <ṯ> /θ/; *Lydian (archaic)* 𐙦, *(classical)* 𐙯 <f>; *Etruscan* 𐙧, 𐙰 <fh> /f/; *Oscan* 𐙧, 𐙱, 𐙲, 𐙳 <f> [f]; *Umbrian* 𐙧 <f> [f]; *S. Picene* 𐙴 <f> /f/. *(i)* Glyph evolution: CT-1, CT-4. *(ii)* Presumably, a /θ/ > /f/ change occurred. The relationship of the S. Semitic <ṯ> /θ/ and the Anatolian and Italic <f> is similar to the relationship of the graphemes with /θ/ or /f/ sound values in SFG-69 and SFG-72. *(iii)* Likely, the Phrygian 𐙯, 𐙦 /b/ph/ also member of SFG-166. (*iv*) The S. Picene 𐙵 <t> (SFG-167) is maybe related to the S. Semitic 𐙦 <ṯ> /θ/.

SFG-167: *Raetic* 𐙶, 𐙷 <t'>; *S. Picene* 𐙷, 𐙸, 𐙹, 𐙵 <t>. *(i)* The Carian 𐙺, 𐙺, 𐙻 <t> /t/ could also be a member of SFG-167. *(ii)* Cf SFG-166.

SFG-168: *P.-Canaanite* (Izbet Sartah, ca. 1100 BC) 𐙼, 𐙽 <ś/š>; *Greek* 𐙾, 𐙿, *Elymian* 𐚀, 𐚁, *Messapic* 𐚀, 𐚂, 𐚃, *Oscan* 𐚄, 𐚅 <σ>; *Lemnian* 𐚆, 𐚇 <ś>; *Lydian* 𐚈, 𐚉, 𐚊 <ś> /s/; *Libyco-Berber* 𐚋 <S3/Ṣ>; *Etruscan* 𐚌, 𐚍, 𐚎 <s/ṣ>; *Phrygian* 𐚏, 𐚐, 𐚑, 𐚒, 𐚓, *SW* 𐚔, 𐚕, 𐚖, *NE-Iberian* 𐚗, 𐚘, 𐚙, *Lycian* 𐚚, *P.-Umbrian* (Tolfa, ca. 530–525 BC) 𐚛, *S. Picene* 𐚜, 𐚝, *Faliscan* 𐚞, 𐚟, *Gallo-Etruscan* 𐚠, *Camunic* 𐚡, *Latin (archaic)* 𐚢, *Lepontic* 𐚣, 𐚤, 𐚥, *Raetic* 𐚦, *Umbrian* 𐚧, *Venetic* 𐚨, 𐚩, *Runic (older fuþark)* 𐚪,

𐤮, 𐤯, 𐤰, 𐤱 <s>; *Madhabic, Sabaic, Hasaitic, Dispersed ONA, Dadanitic, Taymanitic* 𐩦 <s^2> [ɬ]; *Geʿez abjad* 𐩲 <s^2/ś>; *TR* 𐩦, 𐩧, 𐩨, 𐩩, 𐩪 <nč>. *(i)* Cf Old Turkic /š/ ~ /č/ change (Erdal 2004, 103). *(ii)* Glyph evolution: CT-5. *(iii)* Cf SFG-169.

SFG-169: *AH* 𐩦 *380 <sa_8>; *Hismaitic* 𐩧, 𐩦 <s^2> [ɬ]; *Sidetic* 𐩦 <ś> (Adiego 2007e, 14); *SHR* 𐩦, 𐩧, *(cursive)* 𐪉, *CBR* 𐩦 (Table 15 in Appendix), *SR* 𐩦 <s> /s/ (Vékony 1992a, 542), *TR* 𐩦, 𐪊 <s^1, s^2, š, $š^1$, $š^2$> /s, š/. *(i)* The sound values of the Hismaitic <s^2> [ɬ] and of the Rovash <s> /s/ are not identical; however, this is the situation in SFG-168, too. Moreover, the sound value of the Sidetic <ś> has not been clarified. Therefore, the closer known relative of the Rovash is the AH 𐩦 *380 <sa_8>.

SFG-170: *AH* (KARKAMIŠ A31) 𔐲 (Hawkins 2000, 141), 𔐳 (Payne 2010, 14), 𔐴 (Anders 2012), 𔐵 (Hawkins 1986, 371) *402 SCUTELLA <sa_4>, 𔐶 (Payne 2010, 14), 𔐷 (Weeden 2014, 88) *370 <su>, 𔐸 *104 <sà> /s/ [ʃ, s], (MALATYA 6) 𔐹 <us> (Hawkins 2000, 33); *Carian* 𐊧, 𐊨, 𐊩, 𐊪, 𐊫 <ś> /ç?/ (Adiego 2007a, 32, 250; Adiego 2007e, 10) (CT-10); *TR* 𐩪, 𐩫, 𐩬, 𐩭 <$š^1$>, 𐩮, 𐩯 <$š^2$>; *CBR* 𐩭, 𐩮 (Vékony 2004, 164; Hosszú and Zelliger 2014a, 186, 188), *SHR* 𐩮 <š>

SFG-171: *AH* 𐩰𐩱 (Hawkins 1986, 370–371) *389 <tara/i>; *Libyco-Berber* 𐤀𐤀𐤀, 𐤀𐤀𐤀, 𐤀𐤀𐤀, ···· (LBI), 𐤀𐤀𐤀 (Farrujia de la Rosa et al. 2010, 33) <Ṭ/T_1>; *Madhabic, Sabaic (early musnad)* 𐩰, *(early zabūr)* 𐩱, *(middle zabūr)* 𐩲, *(late zabūr)* 𐩳 (Macdonald 2015, 39), *Dispersed ONA* 𐩴, *Taymanitic* 𐩵, *Dadanitic* 𐩶, 𐩷, 𐩸, *Th. D* 𐩹, *Hismaic* 𐩺 (Macdonald 2005, 82), 𐩻, 𐩼, 𐩽 (King 1992, Figure 1 between pages 5 and 6), *Th. B* 𐩼, 𐩾, 𐩻, *Safaitic* 𐩿 (Macdonald 2015, 30, 37), *Hasaitic* 𐪀 <ṭ> /$t^ʕ$/; *Geʿez abjad* 𐪁 <ṭ>

SFG-172: *Lin. A* 𐘀 (Valério 2013, 15–17) 𐘀, 𐘁 LA 37 <ti>; *Lin. B* 𐘁, 𐘀, 𐘁 <ti> /ti, t^hi/; *CM* 𐘁 CM 23 <ti> (Valério 2016, 430–431, 442); *CGk* 𐌕 (Woudhuizen 1984–1985b, 120) *(Common)* 𐌕, (*Paphian,* 6th c. BC) 𐌕 (Olivier 2008, 617–618) <ti> /di, ti/; *Runic (older fuþark, Anglo-Saxon, younger/Danish fuþark)* ᛏ <t> /d, t/

SFG-173: *AH* 𐤊, 𐤋, 𐤌, 𐤍, 𐤎 (Payne 2010, 6, 14, 79, 81) *90 PES <ti>; *CBR* 𐩵, 𐩶, 𐩷 <t> (Vékony 2004, 164; Hosszú and Zelliger 2014a, 188). *(i)* Glyph evolution: CT-10.

SFG-174: *P.-Sinaitic* 𐤩, 𐤪, *P.-Canaanite* 𐤪, 𐤩, *Phoenician* 𐤩, 𐤪, *P.-Hebrew* 𐤪, *Old Aramaic* 𐤪, 𐤪 <t>; *Madhabic, Sabaic* 𐩢; *Hasaitic* 𐩢, *Dispersed ONA, Dumaitic, Taymanitic, Dadanitic, Hismaitic, Th. B, Safaitic, Geʿez abjad* 𐩢, 𐩣 <t>; *SW* 𐩢 <t^a>; *SE-Iberian* 𐩣, 𐩢 <ta>; *NE-Iberian* 𐩢 <ta> /da, ta/; *Celtiberian* 𐩢 <ta>; *Etruscan* 𐌕 <θ> /t^h/; *Raetic* 𐩢, 𐩤, 𐌕, 𐌕, 𐌕 <t>; *Venetic* 𐌕, 𐩢 <t> [d]; *Camunic* 𐩢 <t>; *Umbrian* 𐌕 <t> /t, d/; *Lepontic* 𐩢, 𐩣 <T> /d, t/; *Gallo-Etruscan* 𐩢 <t> /d, t/; *Gallo-Greek* 𐌕, 𐩪 <t> /d, t/; *SHR* 𐌓, 𐌓 <d>; *TR* 𐌓, 𐌕, 𐩪 <d^2>

SFG-175: *Safaitic* 𐩧, 𐩨, 𐩪 (Macdonald 2015, 31, 37) <t>; *SHR* 𐩫 <ť>; *CBR* 𐩬, (Vékony 2004, 192, 197, 198), *SR* 𐩬 <t> (Vékony 2004, 315)

SFG-176: *Greek* (ca. 700 BC) 𐌕, 𐌖 <t> /t/; *Faliscan* 𐌕 <t>; *Umbrian* 𐌕 <t> [d, t]; *SHR* 𐌕 <t> /d, t/. *(i)* Cf SFG-24.

SFG-177: *Greek* 𐌕, *Etruscan, Elymian* 𐌕, 𐌕, *Faliscan* 𐌕, <t, τ> /t/; *Umbrian* 𐌕 <t> [d, t]

SFG-178: *Greek* 𐌕, 𐌕, *Faliscan* 𐌕, *Latin (archaic), Messapic* 𐌕, *Oscan* 𐌕, <t, τ> /t/; *Venetic* 𐌕 <t> /d/

SFG-179: *Greek, Phrygian, Lydian, Lycian, Lemnian, P.-Campanian, P.-Umbrian, Messapic, Elymian, Etruscan, Oscan, Latin (archaic), S. Picene, E. Cyrillic* 𐌕, *Messapic* 𐌕 <t, τ> /t/; *Gallo-Greek* 𐌕, 𐌕, 𐌕, 𐌕, 𐌕, 𐌕, 𐌕 (MNAMON) <t> /d, t/

SFG-180: *Greek* (*late uncial,* 9th c. AD) 𐌕 (Taylor 1883, 154) <τ> /t/; *Glagolitic* (Preslav, ca. AD

893) 𐤕, 𐡕 (NLR) *tvrdo (tverdo)* <t>

SFG-181: *I. Aramaic* 𐡈, *Parthian* 𐡈, *Hatran* 𐣠, *Sogdian* 𐼙 <t>; *TR* 𐤇, 𐤧, 𐤨, 𐤇 <t^2>; *SR* 𐤧 <t^2> (Vékony 1992a, 542)

SFG-182: Writing ductus is boustrophedon in a part of the inscriptions: *AH, CM* (Valério 2016, 179–180, 182, 193), *Latin (archaic), Libyco-Berber, Greek, Hasaitic, Hismaitic, Lemnian, Lepontic, Messapic, NE-Iberian, P.-Sinaitic, S. Picene, Sabaic, Safaitic, SW, Taymanitic, Umbrian, Venetic, TR*

SFG-183: Writing ductus is spiral or circle in a part of the inscriptions: *Etruscan, Latin (archaic), Libyco-Berber, Safaitic, Th. B, Venetic, TR*

SFG-184: Writing versus is bottom-up in a part of the inscriptions: *Libyco-Berber, Safaitic, Th. B*

SFG-185: No word divider in any inscriptions: *Elymian, Hasaitic, Safaitic, Th. B, Th. C, Th. D*

SFG-186: *AH* 𐤈, 𐤈, 𐤈 *216a FINES (ends) *ARHA* <arha>; *SR* 𐤉, 𐤉, 𐤉, 𐤉 separator, end-mark. *(i)* The existence of this SFG is very tentative.

Table 10: Similarity features groups (SFGs)

4.2 Results of the phenetic analysis

The appropriateness of these clusterings strongly depends on the data structure to be clustered. Since the investigated scripts were developed based on a kind of evolution, some branches of the scripts remained close to each other during their evolutionary history. Therefore, the single linkage clustering method is not efficient, since it cannot distinguish clusters with elements close to each other. Moreover, there are outlier members of the script branches, so complete linkage clustering is also not optimal. Certain scripts had several descendants (e.g. Aramaic script), while others remained singular (e.g. Libyco-Berber script). Consequently, the numbers of elements in the clusters largely vary. The UPGMA gives weights to each cluster according to the number of elements of the cluster in each step. Sneath and Sokal demonstrated that the UPGMA would favour clusters more similar in size (fig. 2). Conversely, WPGMA is appropriate when there is a reason *a priori* to eliminate size differences between the resulting clusters. The middle diagram in figure 2 presents the phenogram of the scripts calculated by using WPGMA.

The clearest result is obtained from the Ward method (fig. 2), since it is optimised for homogeneity and filters out the feature similarities that are shared between largely unrelated scripts due to long-term coexistence and cultural interactions. In case of the Ward's method, the square root of Jaccard distance (3) as an Euclidean metric was used.

It is noteworthy that the higher-level joins of the clusters in the dendrograms in figure 2 are analytically uninteresting, since these higher-level joins represent very large dissimilarities in the hierarchical cluster structure. The cluster structure was

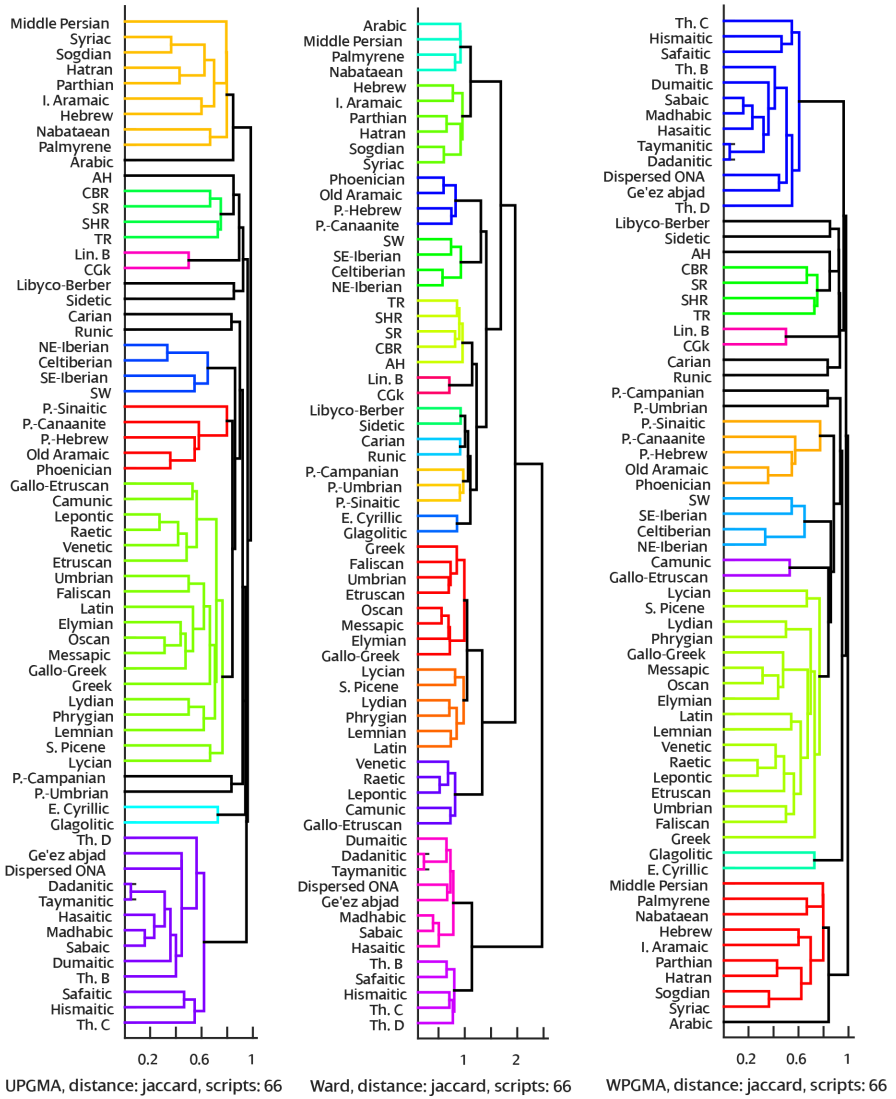


Figure 2: UPGMA (left), WPGMA (middle), and Ward (right) results (M = 66 scripts, N = 186 features)

further refined using leaf ordering methods, which placed leaves next to each other on the dendrogram that are in different clusters but still share some similarity.

Examining the results of the phenetic analysis in figure 2, these mostly medieval, Greek-derivative script Slavic scripts (Glagolitic and E. Cyrillic) were not grouped close to the Greek script. The probable reason for this is that Greek has a large number of glyphs, while the Slavic scripts have much fewer, and thus the calculated distance between them is relatively large. Furthermore, Greek has a large number of other relatives that are unrelated to the Slavic scripts. It can further be observed that the results of all three clustering methods largely agree, differing only in details supporting the stability of the phenetic methods.

An important feature of the k-means clustering is that the mean value of the clusters must be calculated. Consequently, it cannot be used in the case of categorical attributes. Since the features of the scripts can be described with categorical variables, this variable space has to be transformed into a quantitative variable space. For this purpose, multidimensional scaling (MDS) was applied, which transforms the 186-dimensional data points (scripts) to 2- or 3-dimensional data points ($n_{MDS} = 2$ or $n_{MDS} = 3$, respectively). Figure 3 presents the results for the 2-dimensional variable space.

The data points remained representative of the scripts; however, their two quantitative variables (the coordinates in figure 3) are abstract values without interpretable meaning. Then, the k-means clustering was performed on the 2- and 3-dimensional variable space of the MDS output using Squared Euclidean distance; see (4). In the k-means clustering algorithm, the Squared Euclidean distance was used; therefore, each centroid is the mean of the objects in that cluster. The resulting scatter plot in the case of 3-dimensional scaling and $K = 6$ clusters is presented in figure 4; the computation was carried out with the use of MATLAB. The cluster structure was validated by the Dunn index (7), which was 0.7 in the presented case.

The quality of the clusters in figure 4 is measured by the Silhouette index for each cluster, based on (8); see figure 5.

5 Evaluation

5.1 Some observations about the possible origin of some Western Mediterranean scripts

The P.-Hispanic scripts are descendants of Phoenician script. However the phenetic results (SFGs in table 10) present several P.-Hispanic graphemes as being unrelated to the Phoenician. Instead, they are similar to various Aegean, AH, and AGA scripts. The following data support the possibility of transmitting the literacy of the Cypriots to Iberia.

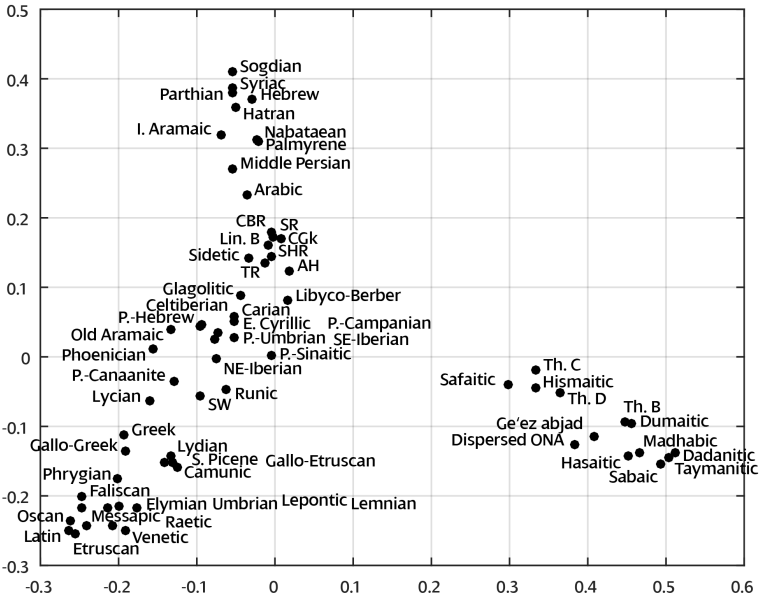


Figure 3: Multidimensional Scaling (MDS), $m = 66$, $n_{MDS} = 2$

According to Botto, between the late 10th and the early 9th c. BC, the Phoenicians used their strongest bond, with the Cypriot element, to penetrate the southern Tyrrhenian Italian and Sardinian markets. Moreover, in the late Bronze Age connections existed between southern Iberia and Sardinia. There was an alliance between the main Phoenician and Cypriot coastal cities. In the 11th–10th c. BC between Cyprus and Sardinia, the relationships became vital. After the fall of Mycenaean power, the Cypriots played a significant role in trade between Levante and the western part of the Mediterranean (Botto 2016).

Another important fact is that the P.-Hispanic scripts are syllabic for the plosives and alphabetic (monophonemic) for the rest of the consonants and the vowels. Moreover, the syllabic graphemes for the plosives do not mark a voicing contract. This is the reason why they are called semi-syllabaries. Especially interesting is the so-called *principle of redundancy,* which means that in one of the P.-Hispanic scripts, the SW, each syllabic grapheme is accompanied by a redundant grapheme representing the vowel of the syllabic grapheme (Valério 2008, 112; 2014, 440). A possibly related fact is

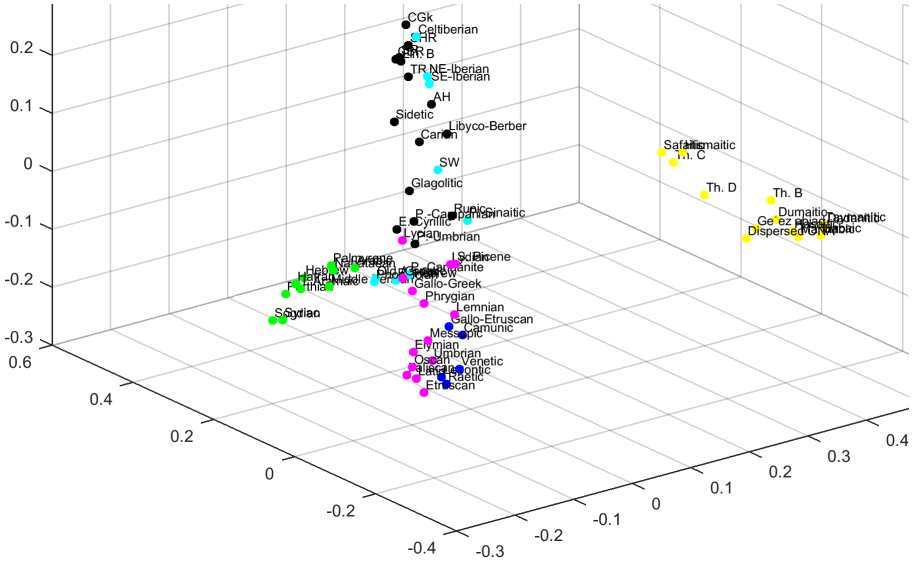


Figure 4: K-means mapped to MDS, $m = 66$, $n_{MDS} = 3$, $K = 6$

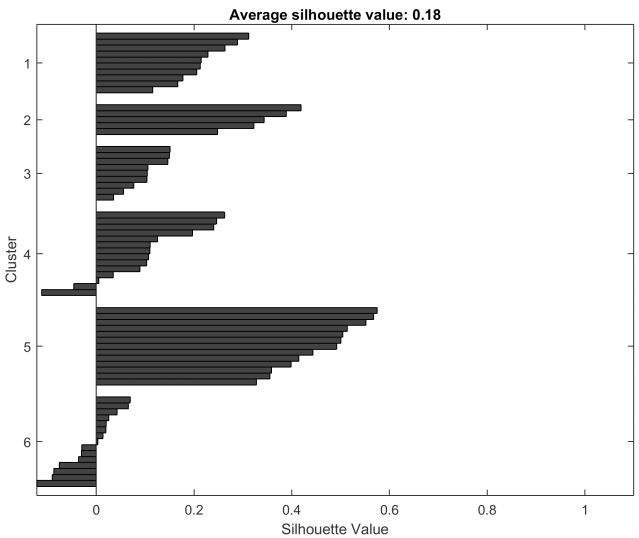


Figure 5: Silhouette values of k-means clusters

that in the Assyrian cuneiform and the AH scripts, word-ending long consonants are represented with the <CV> + <V> grapheme combination, e.g. <ki>+<i> represented /kī/ (Segert 1978, 111–112). Consequently, it is not ruled out that the P.-Hispanic scripts were affected by the AH. This conforms to Valério's supposition—citing Craig Melchert—that the AH 𐔎 *376 <zi> at least indirectly affected the SW script (Valério 2008, 130–131).

Considering the Ancient Italic scripts, the northern version of the Etruscan script probably originated from Lydia, and the southern version could be from Cilicia (Woudhuizen 1982–1983, 98). Woudhuizen claimed that in Pithecussae there was the presence of Lydians in the 8th c., who disappeared in the 7th c. BC. According to Woudhuizen, Lydian and eastern Greek merchants founded Pithecussae in the early 8th c. and Cumae in the late 8th c. According to Szabó (2015, 352), the Etruscan territories around Bologna were affected by northern Balkan and Hallstatt archaeological features, and oppositely, the southern Etruscan areas were influenced by Anatolian features. These archaeological data are used only in support of the results presented here, and no archaeological conclusions are drawn. However, a possible consequence of these (and several others, not cited) archaeological data could be that the culture in Italy was heterogeneous, which could lead to the preservation of ancient glyphs in the orthographies. Considering the dendrogram obtained by the Ward method in figure 2, it is interesting to note the strong relationship between the Runic, some Ancient Italic, and AGA scripts. It could imply that the spread of writing knowledge in Italy happened in multiple waves. In such case, the Runic maybe preserved an early layer of literacy in Italy. This approach is not contradictive to the model of Looijenga (1997, 55–56), who demonstrated that the Runic originated by adaptation of some kind of Northern Italic local script in Romanized regions along the Rhine.

To summarize, besides the Phoenician, there was another source of the P.-Hispanic, Libyco-Berber, Ancient Italic, and Runic scripts that could be rooted in the eastern Mediterranean.

5.2 An approximative model for the origin of the Rovash scripts

Based on the SFGs in table 10, there are several graphemes that are very similar to the Rovash graphemes. The possible donors or close indirect relatives of the Rovash graphemes are summarized in table 11 and table 13 with SFG references to table 10. The graphemes of scripts that might have been earliest affected by the Rovash script are listed in table 11. Considering the very early age of use of the Lin. A, Lin. B, and CM scripts (table 4), if similar glyphs exist in different Aegean scripts, only the latest occurrence is denoted in table 11, usually in the CGk script. However, in some cases, the most similar glyph occurred in earlier scripts than the CGk. The Rovash 𐔈, 𐔉 <d, d^2> grapheme was left out from table 11, since it is not possible to choose the most probable source due to the large number of candidates in SFG-174.

Group	Script	Probable donor or indirect relative of the Rovash glyph	Sum
	CM	CM0 09 <ka?> (SFG-87); CM 112 <k/ze?> (SFG-92, or LA 74 <ze?>); CM , 70 <ki?> (SFG-97, cf Lin. B , <ki>, CGk , <ki>); CM 75 <ra> (SFG-160, cf CGk <ra>); CM 24 <le?> (SFG-162, cf CGk <le>)	16
Aegean	CGk	<za?> (SFG-54); <ja> (SFG-75); <ka> (SFG-88); <me> (SFG-116); <mi> (SFG-118); <mo> (SFG-122); <mu> (SFG-123); <ne> (SFG-133); <sa> (SFG-141); , , <se> (SFG-145); <pa> (SFG-148, SFG-149)	
Anatolian hieroglyphic	AH	*451 <hur> (SFG-31); , , *376 <zi> (SFG-60); *315 <kar> (SFG-90); *423 <ku> (SFG-100, or CM 15 <ko?>); *35 <na> (SFG-132); *411 <ni> (SFG-135); , *383 <ra/i> (SFG-161); , *380 <sa_8> (SFG-169); *402 <sa_4>, <us> (SFG-170); <ti> (SFG-173); *216a FINES (SFG-186)	11
Anatolian-Greek alphabetic (AGA)	Carian	<δ> (SFG-29); <γ> (SFG-101); <d> (SFG-30)	17
	Greek	<α> (SFG-7); <β> (SFG-13); <ε> (SFG-33); <ϝ> (SFG-42); <ζ> (SFG-52); , , , <θ> (SFG-69, or Old Aramaic , , <ṭ/ṯ>); <l> (SFG-108); <λ> (SFG-109); , <σ> (SFG-168); <τ> (SFG-176)	
	Lycian	<g> (SFG-18); <q> (SFG-103, or CGk <ku>)	
	Sidetic	<a> (SFG-12, cf AH *19 <á>); <g> (SFG-98, or Greek <kh?>)	
Ancient Semitic & Canaanite	Proto-Sinaitic	, <h> (SFG-31); , , <ḫ/ḥ> (SFG-68)	6
	Phoenician	<h> (SFG-33); <m> (SFG-112, or Old Aramaic <m>); <n> (SFG-124)	
	Old Aramaic	<r> (SFG-158)	
S. Semitic	S. Semitic	<'> (SFG-4, cf Phoenician <'>); , , , <h> (SFG-32); , <y> (SFG-47); , <ṯ> (SFG-57); <g> (SFG-102); <s^1> (SFG-143, cf CGk <sa> in SFG-141); <s^1> (SFG-144, cf *AH* *415 <sa>, CGk <sa>); <r> (SFG-164, cf CGk <re> and Sidetic <r> in SFG-163); <t> (SFG-175)	9

Table 11: Presumably direct donors or close indirect relatives of the Rovash graphemes

The SFGs suggest that the common ancestor of the Rovash scripts had to have been developed in Anatolia, after the distribution of the Semitic consonantal scripts (Proto-Sinaitic, Phoenician, Old Aramaic), likely before the end of the syllabaries, which originated from the AH and the Aegean scripts, and surely before the 3rd c. BC, when the Greek script became dominant in Anatolia.

It is known that the Turkic Rovash (TR) originated from the nomad region of the Altai Mountains (Vasil'ev 1994, 328). From this it follows that a presumably common ancestor of the Rovash scripts (Proto-Rovash) must have reached the Altai Mountains beforehand. According to Marsadolov (2000a, 247–250; 2000b, 51), during the 6th – 4th c. BC the Pazyryk *(Pazîrîk, Пазырык)* culture ruled the Altai region, and the descendants of the Cimmerians, who settled there after being expelled from Anatolia, may have participated in the growth of the Pazyryk culture, which was also influenced by the Chinese and Achaemenid Persian empire. In the first half of the 6th c. BC, numerous innovations appeared in the Altai region which, according to Marsadolov, can be linked to the arrival of nomadic tribes from Anatolia at the end of the 7th c. BC or the beginning of the 6th c. BC. Presumably the nomadic tribes from Gordion or the surrounding region settled the most fertile valleys, Tuekta and Bashadar, assuming rule over the local Pazyryk population. The 4000 km distance between Anatolia and the Altai region was not insurmountable, in part due to the existence of trade routes. The nomads could have crossed this distance in as little as one to two years.

The Altai region later became part of the Yüeh-chih *(Yuèzhī)* Empire, and the sites at Pazyryk should be related to the Yüeh-chih (Enoki et al. 1994). According to Harmatta, the Yüeh-chih is known as Tochari in Greek and Latin sources. Between 203 and 177/176 BC, the Hsiung-nu *(Xiongnu)* defeated the Yüeh-chih, who migrated to the west (Harmatta 1994).

In the necessary timeframe (7th–6th c. BC), there is knowledge of only a single ethnic group, the Cimmerians, who could have taken literacy from Anatolia to the East. The Cimmerians seized Phrygia from King Midas in the first half of the 7th c. BC. During the same period, Caria fell to the Lydians (Adiego 2007b, 758). For generations the Cimmerians lived around Gordion (the late Phrygian capital), making two attempts to capture Lydia during 650–640 BC. Eventually the Lydian ruler Alyattes expelled the Cimmerians from Anatolia in the late 7th – early 6th c. BC (Marsadolov 2000a, 249).

If the Cimmerians borrowed the later Rovash graphemes, the S. Semitic scripts could not be ancestors of these graphemes, since they had no known contact. More probably, both the Rovash and S. Semitic scripts originated from a common region; see also table 12. Based on the SFGs, it seems very likely that there are numerous S. Semitic graphemes of *non-Phoenician origin.* Moreover, these non-Phoenician S. Semitic glyphs appear in other scripts of the first half of the 1st millennium BC. Table 12 presents each presumably non-Phoenician S. Semitic grapheme and the occurrence

S. Semitic	Ancient Italic	Aegean	AH	AGA	Libyco-Berber	P.-Hispanic	Rovash	Runic
𐤀 <’> (SFG-3)	0	0	0	0	0	0	1	0
𐤉 <’> (SFG-32)	0	0	0	0	0	0	1	0
𐤉 <y> (SFG-47)	0	0	0	1	0	0	1	0
𐤆 <z> (SFG-53)	0	0	0	0	0	0	1	0
𐤇 <ḍ> (SFG-70 and SFG-71)	0	0	0	0	0	1	0	0
𐤂 <g> (SFG-87)	1	1	0	1	0	1	1	1
𐤂, 𐤂 <g> (SFG-100)	1	1	1	1	0	1	1	1
𐤂 <g> (SFG-101)	0	1	0	1	0	1	1	0
𐤂 <g> (SFG-102)	0	0	0	0	0	0	1	0
𐤋 <l> (SFG-107)	1	0	0	1	0	1	0	1
𐤌 <m> (SFG-118)	0	1	0	1	0	0	1	0
𐤍 <n> (SFG-127)	1	0	0	1	0	1	0	0
𐤍 <n> (SFG-128)	1	0	0	0	0	0	0	0
𐤎 <s^1> (SFG-143)	0	1	0	0	0	0	1	0
𐤎 <s^1> (SFG-144)	0	1	1	0	0	0	1	0
𐤓, 𐤓, 𐤓 <r> (SFG-164)	0	0	0	0	0	0	1	0
𐤔, 𐤔 <t̲>, 𐤔, 𐤔 <f> (SFG-166)	1	0	0	1	0	0	0	0
𐤔 <s^2> (SFG-168)	1	0	0	1	1	1	1	1
𐤔 <s^2> (SFG-169)	0	0	1	0	0	0	1	0
𐤕, 𐤕, 𐤕 <t> (SFG-175)	0	0	0	0	0	0	1	0
boustrophedon (SFG-182)	1	1	1	1	1	1	1	1
spiral or circle (SFG-183)	1	0	0	0	1	0	1	0
bottom-up (SFG-184)	0	0	0	0	1	0	0	0
Summary	9	7	4	10	4	8	17	5

Table 12: Occurrence of cognates of non-Phoenician S. Semitic graphemes

of their counter pairs in other scripts based on the SFGs in table 10. For this study only, SFG-70 and SFG-71 were hesitantly unified.

The resulting numbers of cognate graphemes in table 12 cannot be evaluated quantitatively, since the populations of each group of the examined scripts are largely different. Nevertheless, it can be observed that scripts other than Rovash had a significantly weaker relationship with the S. Semitic scripts. Consequently, the region where graphemes were transferred to a supposed ancestor of the Rovash scripts was

presumably not farther from the region from which the S. Semitic scripts borrowed certain graphemes than was the region that lent graphemes to other examined scripts. Based on the known historical data, the S. Semitic groups did not reach any region in Anatolia except a part of the Neo-Hittite states (Syria and Southeastern Anatolia). Therefore, the region lending graphemes to Rovash scripts could not be far from the Neo-Hittite states.

According to Macdonald, two forms of the Sabaic script (a kind of S. Semitic) are the formal musnad and the informal, cursive zabūr. Several zabūr relics have been carbon dated and found that the oldest one was from the period 1150-901 BC with a confidence of 2σ (94%) (Macdonald 2009, Addenda and Corrigenda, 10). Consequently, the common ancestor of the S. Semitic and Rovash scripts could not have developed later than the 11th c. Since the start of the CGk is about the 11th c. BC (Valério 2016, 237), this may justify that the most similar Rovash graphemes may have come from the CM script, which was still used in the 11th c. BC (Valério 2016, 27), rather than the CGk (table 11).

Lehmann claims that in the 12th-11th c. BC, both Syria and Cilicia were affected by the Aegean culture in part due to the Aegean settlers in the costal regions and also due to the Aegeans' trade with Syria, Lebanon, Cyprus, and Cilicia at the end of the Bronze Age (Lehmann 2013, 265, 325, 328). According to Yakubovich, the Cilician leaders were of Greek-speaking Aegean descent in the Early Iron Age (Yakubovich 2015b, 35–36, 38, 40–41). Thus, the Cypriot scripts could have affected Cilician literacy.

Que (Assyrian name; its Luwian form was Hiyawa) situated on the Cilician plain was one of the Neo-Hittite states (Yakubovich 2015b, 49). Greek pottery from the 12th–11th c. BC is found in large quantities in the Cilician plain. The Greek settlers in Pamphylia succeeded in establishing their linguistic dominance in this region. Cilicia represents the only region where Luwians and Greeks may have coexisted. A neighbour of the Greeks in Southwestern Anatolia was the Carians (Yakubovich 2008, 200). The main official language of Que was not Luwian, even though Luwian was historically spoken by the bulk of its population. The socially dominant language was Greek, and the attested written language is Phoenician.

According to Yakubovich, the Phoenician language was emblematic of the rulers of Que, who claimed Greek descent, and the Luwian language was used by the indigenous population of Que from before the collapse of the Hattusa empire. Yakubovich claims that the adoption of Phoenician as a language of written expression by the Greek colonists in Cilicia happened at the point when the Linear B script had been forgotten and represented the first step toward the creation of the Greek script. Furthermore, the Greek script originated from Cilicia in the late 9th century BC. In Que, no Semitic personal names are attested to in these inscriptions in connection with local individuals. Valério (2008, 116) claims that the Phoenician script was used for recording Luwian personal names. Swiggers (1996, 266–267) stated that the Cilicians

Group	Invoked graphemes of the donors	Sum
Aramaic & Persian	I. Aramaic 𐡅 <w> (SFG-44); I. Aramaic 𐡉, 𐡊 <y> (SFG-81); I. Aramaic 𐡐 <p> (SFG-151); I. Aramaic 𐡕 <t> (SFG-181); Sogdian orthographical rule <w>+<y> for representing /ö, ü/ (uncertain, see comments in SFG-45); Middle Persian 𐭩 <y> (SFG-83, or Palmyrene 𐡉 <y>);	6
Slavic	Glagolitic: ⰻ <o> (SFG-146); E. Cyrillic: Ѫ <ǫ> (SFG-147)	2

Table 13: Sporadic influence on the Rovash scripts

could have adopted the Phoenician script but only used it for the inscriptions in the Phoenician languages. Thus it is proven that the Phoenician script was present in Cilicia.

Although no local script relics in the Greek language have been found for the relevant place and period, several new findings and methodological advances made since the year 2000 have strengthened the case for a Greek existence in Early Iron Age Cilicia (Yakubovich 2015b, 49). The Early Iron Age assemblages excavated in Cilicia match those of the northern Levant in attesting to the presence of materials connected with the Aegeans. From the period between the 12th to mid-8th c. BC, no AH inscription was found in Cilicia. Cilicia is the only region of south-central Anatolia and northern Syria in which a Neo-Hittite tradition begins in the very late 8th century BC without any earlier trace of a post-Hittite tradition (d'Alfonso and Payne 2016). Consequently the AH script did not dominate in Cilicia.

Based on the above data and geographical factors (the Cimmerians were neighboured by Cilicia), it is likely that the Rovash graphemes originated from the region around Cilicia. Similarly, non-Phoenician S. Semitic glyphs in table 12 may also have originated from the Cilicia region.

In the period 700–200 BC, the I. Aramaic, from 1st c. BC to 7th c. AD, the Late Aramaic, and the Middle Iranian in Central Asia, and around 10th c. AD, the Slavic scripts in the Carpathian Basin affected the Rovash scripts by the graphemes listed in table 13. It is noteworthy that some of the I. Aramaic graphemes (e.g., 𐡅 <w>, SFG-44) could have been borrowed in the earliest time, maybe even in Anatolia.

The Rovash ᚹ, ᚻ, ᚼ, ᛬ <W> (SFG-44) may have originated from the Old Aramaic 𐤅, 𐤆 <w> or the I. Aramaic 𐡅, 𐡆 <w>, and surely not from the Old Aramaic 𐤕, 𐤄, 𐤅, 𐤇, 𐤉 <w> (SFG-41). In the case of the Rovash <i, y>, the typical glyphs are ᚻ, ᛁ, ᛦ (SFG-81), which are unrelated to the Phoenician 𐤆, Old Aramaic 𐡆, and I. Aramaic (7th c. BC) 𐡆, (6th c. BC) 𐡆, 𐡇 <y> (SFG-77), but more probably related to the Old Aramaic (7th c. BC) 𐤊, 𐤋, (6th c. BC) 𐤊, 𐤊, 𐤋, 𐤉 <y> (SFG-81). Consequently, the adaptation of

TR grapheme	Supposed meaning in Old Turkic	SFG with some example glyphs
𐰉, 𐰊 <b^2>	*äb, äβ* 'tent, house'	SFG-116, e.g., P.-Hispanic 𐰉 <be>
𐰋 <d^2>	*ed* 'property, livestock'	SFG-174, e.g., P.-Hispanic 𐰋 <ta> /da, ta/, Ancient Italic 𐰋, 𐌕 <t, T> /d, t/
𐰍, 𐰏, 𐰑 <g^1>	*aγ* 'net'	SFG-103, e.g., Lycian 𐊧, 𐊨, 𐊩 <q>
𐰃, 𐰆, 𐰇 <y^1>	*ay* 'moon'	SFG-75, e.g., CGk 𐰇, 𐰃, 𐰆 <ja>
𐰴, 𐰵 <k^3/^{ï}k^{ï}>	*ïq* 'spindle'	SFG-97, e.g., Carian 𐊫, 𐊬, 𐊭 <k>
𐰚, 𐰞 <k^5, ^{w}k^{w}>	*oq* 'arrow'	SFG-88, e.g., CGk 𐰞 <ka> /ga, ka, k^{h}a/
𐰽 <l^2>	*el* 'hand'	SFG-141, e.g., CGk 𐰽 <sa> (see comments in SFG-141)
𐰣, 𐰤 <n^2>	*en* 'declivity'	SFG-133, e.g., CM 𐊲, 𐊳 34 <ne?>; CGk 𐰣, 𐰤 <ne>
𐰼 <r^2>	*er* 'man'	SFG-165, e.g., Lin. B 𐰼 <ru> /lu, ru/, Lydian 𐤚, 𐤛, 𐤜 <λ>
𐰀, 𐰁, 𐰂, 𐰃, 𐰄, 𐰅, 𐰆, 𐰇, 𐰈, 𐰉 <t^1>	*at* 'horse'	SFG-29, e.g., Carian 𐊠, 𐊡 <δ> /md/d/^{n}t/, SW 𐊢 <t^{o}>

Table 14: The relationship of the TR graphemes that were traditionally supposed to be ideograms

the Aramaic <y> may have happened in the 7th– 6th c. It is noteworthy that there is more cursive Rovash 𐰃 <i, y> (SFG-83), which had to have been adapted in the period 1st c. BC – 7th c. AD. Due to historical reasons, this adaptation had to have happened in Middle Asia. In the case of the Rovash 𐰼, 𐰺, 𐰽, 𐰾 <r>, the ancestor is surely the Aramaic 𐡓 <r> (SFG-158). The strictly geometric forms of the Rovash glyphs point to an early adaptation; however, cf Turkic Rovash 𐰺, 𐰽 <r^1> and I. Aramaic (Aśoka, ca. 250 BC) 𐨪 <r>.

5.3 The question of the TR ideograms

Several authors have hypothesized that some of the TR graphemes originate from ideograms (pictograph, tamgha). The history of this direction of research is summarized in Róna-Tas (1987, 8). However, similar counterpairs of the TR graphemes in question (Róna-Tas 1987, 9) can be found in the SFGs of table 10, as is demonstrated in table 14. The listed example glyphs in the last column of table 14 are usually not direct relatives of the appropriate TR graphemes; however, they show the probable relationships of the TR graphemes in question. Using the *lex parsimoniae,* it is unnecessary to assume they have an ideogrammatic origin.

5.4 Syllabic traces in the Rovash scripts

There are traces of syllabary in the Turkic Rovash; namely, Kyzlasov (1994, 131) explored that the TR is partly a syllabary. Kyzlasov claimed that the ancestor of the TR (he called: проторуническое слоговое письмо) goes back not to alphabetic systems but to the ancient, probably Semitic, syllabaries of an unknown (not West Semitic) origin. He claimed that the ancestor script developed by eliminating a part of the earlier used presumed syllabic graphemes, and the surviving Orkhon and Yenisei inscriptions demonstrate the final stage of this process. According to Kyzlasov, the ancestor of the TR was not invented but borrowed. He further supposed that among many such systems the ancient Turkic "linguists" wisely chose the alphabetical system best suited to the Turkic language. The outward similarity of the symbols of the various Euro Asiatic and Asiatic Turkic inscriptions can be explained by their basis, the ancient Semitic scripts of Central Asia. Each of these versions of writing systems used for TR inscriptions was formed under different conditions and on a different basis. In the reconstruction by Kyzlasov, most of the consonants are denoted by two different kinds of graphemes, depending on the vowel in the syllable of the consonant (velar or palatal sound values). A consonant is called *velar* if it is used near back vowels, and it is called *palatal* if it is used near front vowels. Consonants are harmonized with the vowels of their syllables. Graphemes that represent consonants next to back and front vowels are transliterated by adding a superscript 1 or 2, respectively, to the transliteration value of the consonant, e.g. b^1 and b^2.

As Erdal (2004, 39) pointed out, synharmonism (vowel harmony) and the presence of the front rounded vowels *ö* and *ü*, both are equally untypical of Semitic, Caucasian, East Asian, and Early Indo-European. The TR script distinguishes front and back harmony in rounded vowels and also in consonants; there are, for example, sets of very different-looking graphemes for front *b* and back *b,* front *y* and back *y,* etc. The palatal consonant *y* is sometimes used in the Old Turkic language beside front vowels. Semitic scripts distinguish only between velar and uvular /k/ (*'k'* and *'q'*) and /g/ (often noted *g* and *γ* respectively), a distinction which has been used for expressing synharmonism in Turkic languages. A further specific feature of TR is the preponderance of closed syllables as opposed to open ones. For example, unlike Semitic and Indo-European scripts, the grapheme for a consonant *t* implies not a following vowel, but a preceding vowel. Moreover, all coda vowels are written out as separate features in the TR, again unlike the Semitic and Indic scripts (Erdal 2004, 39–40). Possibly related to Erdal's observation is that in the earliest SHR relics, the consonant grapheme names begin with a vowel (to ease pronunciation), different from the usual European practice where the vowel is placed after the consonant.

Following are known synharmonism of consonants in the TR: /b/, /p/, /d/, /t/, /g, ɣ/, /k, q/, /t/, /l/, /n/ /r/, /s/, /j/, and /ïq/, /oq/, /üq/. Moreover, as Kyzlasov claimed,

certain graphemes could have been used for syllables /ït/, /ïš/, /ïs/, /id/ed/, /ič/eč/, /im/em/. His reconstruction supports the possibility that the common ancestor of the Rovash scripts originated from at least partly syllabaries. However, no known Rovash script is a syllabary. Even synharmonism exists only in the TR. Sporadically and not consequently, the SR also applied synharmonism in the case of some consonants, as Vékony (2004) demonstrated. In the CBR and SHR there are some consonants with multiple graphemes: In the CBR, for the /k/ and /t/, and in the SHR, for the /č/, /k/, /š/, /r/, and /t/, there are multiple graphemes; the reason for this has not been clarified. In the surviving CBR and SHR relics, usually there is no synharmonism. However, in a very few SHR relics, the differentiation of the <k> graphemes near front and back vowels can be detected. Moreover, in the Constantinople inscription, the grapheme ᛋ <k> seems to represent also the syllable /a:k/ besides the consonant /k/ (table 17 and comments).

Consequently, the Rovash scripts may have preserved traces of an ancient syllabary, but there is no evidence for an ancient syllabary as the common origin of the Rovash scripts (Proto-Rovash). However, taking into account the fact that, according to the phenetic analysis, several graphemes of the semi-syllabic P.-Hispanic and the Rovash scripts (see SFGs in table 10) are markedly similar, it can be supposed that the Proto-Rovash could have had some syllabic property.

5.5 Witness scripts as a consequence of the centre-periphery effect

In the 3rd–2nd millennia BC, the centre of script development was in the Middle East. Presumably, the North-West Semitic and the S. Semitic writing traditions separated in the 2nd millennium BC (Macdonald 2015, 32). In the 1st millennium BC, it gradually diverged into the Aramaic world (east) and the Anatolian-Greek world, and later (classical) Italy. Using Macdonald's model for literate and non-literate societies, in these areas the societies were literate; therefore, these places can be considered central. Conversely, in the nomadic or partly nomadic Arabian Peninsula, Hispania, Northern Africa, and the Eurasian Steppe from the eastern Altai Mountains to the western Carpathian Basin, the societies can be modelled predominantly as non-literate; therefore, they are considered peripheral. Theoretically, the peripheries could preserve glyphs that were already forgotten in the centre in favour of the later developments. A centre–periphery (core–periphery) model can be used for the spatial distribution of certain glyphs.

The extracted SFGs (table 10) clearly show the significant similarities in several glyphs and orthographical rules in the Ancient Italic, Libyco-Berber, P. Hispanic, Rovash, Runic, and S. Semitic scripts. Taking into account some historical facts, these scripts probably originated from Levantine or the Anatolian coast. All of these scripts left Anatolia not later than the 7th c. BC; therefore, they could have preserved a

certain state of the grapheme evolution in Anatolia. These groups of scripts can be qualified as witnesses of the graphemes used in Anatolia and the surrounding regions in the first half of the 1st millennium BC.

It is noteworthy that the property of being a witness script is a relative quality, since a certain script could be witness of the development of another script, which could be witness of another. For example, the AGA scripts also witness the age of their development; they testify a mainly alphabetic environment from the early age of the Greek script. The beginning of the AGA scripts is about the 8th c. BC (table 4), based on the earliest dated inscriptions. The accurate development of these scripts remains unknown, however, they did preserve even earlier graphemes, such as the Lycian ◊ <k> (SFG-100).

6 Conclusions

The paper presented a new composite data analysis method to explore the similarities between scripts. Computational palaeography concentrates on the topological relationships of each grapheme. The premise is that the glyphs of the graphemes are relatively stable during the development of the writing, and the changes can usually be described by well-defined rules. During this, the linguistic, historical, geographical, and archaeological circumstances are taken into account as accurately as possible.

The developed method starts with searching for sets of possible cognate glyphs. It utilizes the determined typical *characteristic transformations* of the topology of the glyphs, which can be observed on the evolution of the graphemes. The characteristic transformation usually does not change the visual identity of the original glyph. The topological and the visual identity layers belong to the layered grapheme model, which was developed for modelling the grapheme in computational palaeography. The developed data analysis method selects orthographical rules and sets of possible cognate glyphs from the phonetically similar graphemes by minimizing the necessary topological transformations between glyphs. In such way, the similarity features groups are constructed. Then various machine-learning methods are applied to obtain a phenetic model for the investigated scripts based on the similarity group of features. In this stage, the multidimensional scaling and various clustering algorithms were applied. The obtained results give an overall picture about the phenetic relationships of the examined scripts. In order to filter out the possible homoplasies, a cladistic approach was also used, in a limited fashion.

Some special concepts were elaborated and introduced in the computational palaeography in order to apply the phylogenetic methods for palaeography. Beside the existing term of characteristic transformation, the concept of the witness script is also introduced. A script is taken as a *witness script* for a certain area and time period

if the continued evolution of the script happened in isolation. Further new concepts are the glyph- and grapheme-level reticulations as reticulate events. A *glyph-level reticulation* occurs if part of the glyphs of a grapheme is borrowed from another script, and a *grapheme-level reticulation* exists if all glyphs of a grapheme are borrowed in a certain evolutionary event.

The results show the usability of the phenetic approach combined with cladistic elements in exploring the similarities of scripts. The present study concentrated on the phenetic analysis of Mediterranean-origin scripts; but the presented method could be extended to other writing systems. The main goal was to prove the usability of the combined exploratory data analysis method; however, during the evaluation of the resulting phenetic model, some approximative consequences can be derived about the relationships of the examined scripts as follows. *(i)* Some groups of witness scripts are identified which attest the state of the grapheme evolution in the first centuries of the Iron Age in the Mediterranean. These are the S. Semitic, the P.-Hispanic, the Ancient Italic, the Libyco-Berber, the Runic, and the Rovash. *(ii)* The origin of these witness scripts is at least partly connected to south Anatolia. *(iii)* The probable source of the Rovash graphemes was approximately determined as the region of Cilicia before the 6^{th} c. BC.

The developed method for script analysis might be used for further applications. Changing the focus of the research, it is possible that the basic taxonomical unit (taxon) is not the script, but a version of the script (e.g., grapheme set of the medieval English orthography), or a certain writing style, typography, and so on. The introduced approach may give support to palaeographers in exploring the relationships among scripts and deciphering ancient inscriptions. The present method can be highly automatized; therefore, it could be scaled to library-wide databases.

7 Appendix: Examples of Rovash inscriptions

7.1 A quadrilingual CBR inscription of the Golden Treasure of Nagyszentmiklós

The Golden Treasure of Nagyszentmiklós is a tableware collection of 23 gold pieces found in Nagyszentmiklós, Hungary (currently Sânnicolau Mare, Romania) on 3 July 1799. The treasure is unique in the region; the total weight is 10 kg. Its style cannot be connected to any great cultural center; most probably, it is a local product, made in the 7^{th}–8^{th} c. AD (Bálint 2010); however, the majority of the inscriptions could have been carved later. The names of the beverages to fill the jugs and the names of the foods to be served on the plates were carved onto the bottom of the pieces in CBR script (Vékony 2004). That is why the Rovash texts are mainly names of drinks and food.

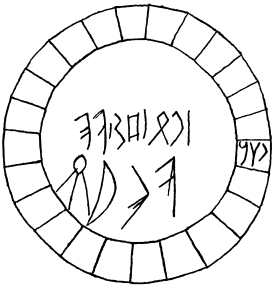


Figure 6: Drawing of the quadrilingual Nagyszentmiklós inscription (Hampel 1884)

First row in the middle	/sïu̯ *s^{o}r^{i}m/ *(right part in Ogur),* /β^{i}z^{i}/ *(left part in Hungarian)* 'water [and] beer/wine' *(right part),* 'water' *(left part)*
Second row in the middle	/v^{o}d^{o}jõ/ *(in Slavic)* 'with water'
Text in the sidewall frame	/dan$^{(u)}$/ *(in Alan)* 'water'

Table 15: The transliteration, transcription, and translation of the quadrilingual Nagyszentmiklós inscription

One piece of the treasure, jug No. 6, is a unique relic, since the inscriptions on its underside (fig. 6) are in different languages having similar meaning (table 15). In the first row there is a short vertical bar, which separates the Ogur (Turkic) text (right) and the Hungarian text (left). The detailed palaeographical analysis of this inscription, including the alternative readings, is published in Hosszú and Zelliger (2014a).

7.2 The Vargyas SHR inscription

A Székely-Hungarian Rovash inscription was found in Vargyas (Romanian: Vârghiș, Székelyland, Romania) on a stone in a church in 1994 (Benkő 1994, 487–489). For linguistic reasons the text could not have been created later than the 12th–13th c. AD (Zelliger 2016). The drawing of the inscription is presented in figure 7. The writing of the inscription is sinistrograde; see table 16. The sentence in the inscription was cited from the Gospel of John (Ioh 19:26): "*Woman, behold your son.*" Zelliger claimed that

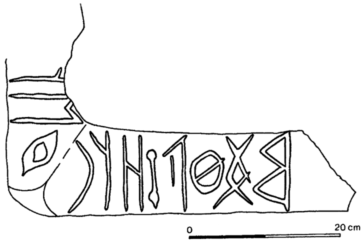


Figure 7: Drawing of the Vargyas inscription (Benkő 1996a, 79; 1996b, 31–33)

Transliteration with Rovash graphemes	ᛘᛖ◊Ↄ ᛘ ᚺᛧᛏ0 ⨯ᐊ
Transcription with phonetic symbols	/$^{\text{i}}$mē fioɣ t$^{\text{e}}$ n$^{\text{ä}}$küd/
Translation to English	‘[Woman,] behold your Son’

Table 16: The transliteration, transcription, and translation of the Vargyas inscription

the SHR text of the Vargyas Inscription resembles a Greek translation of the Bible. The detailed palaeographical analysis of this inscription, including the alternative readings, is published both in Zelliger (2016) and in Hosszú (2013).

7.3 The Constantinople SHR inscription

In 1515 in Constantinople (Istanbul), Barnabas Bélay, the ambassador of the Hungarian King Vladislaus II (1490–1516), found he had to wait for two years for his admittance to the Sultan Selim I (1512–1520), and during this time, a Hungarian person named Thomas Kidei Székely wrote this SHR inscription on the wall of the Ambassadors' House. Between 1553 and 1555, the numismatist and epigraphist Hans Dernschwam (1494–1568 or 1569) discovered and copied it (fig. 8); later the building was destroyed in an accidental fire (Babinger 1914, Sebestyén 1915). The writing of the inscription is dextrograde; see table 17. The detailed palaeographical analysis of this inscription, including the alternative readings, is published in Zelliger and Hosszú (2014).

The inscription contains several ligatures, e.g., the symbol Ħ (first row) is maybe the ligature of the graphemes *F <e> (SFG-33) and *H <r> (SFG-158), the sound value of the ligature being /er/. The glyph *F is presumably the mirrored version of the SHR ᑣ <e>, which is attested in the Dálnok and the Rugonfalva inscriptions (Hosszú and Zelliger 2013).

In the Constantinople inscription, the ᛦ <k> is used consequently in the syllables containing /a, ā/ vowels. Therefore, it is possible that the sound value of ᛦ <k> was /ak, aːk/ in the orthography of the Constantinople inscription. This is supported

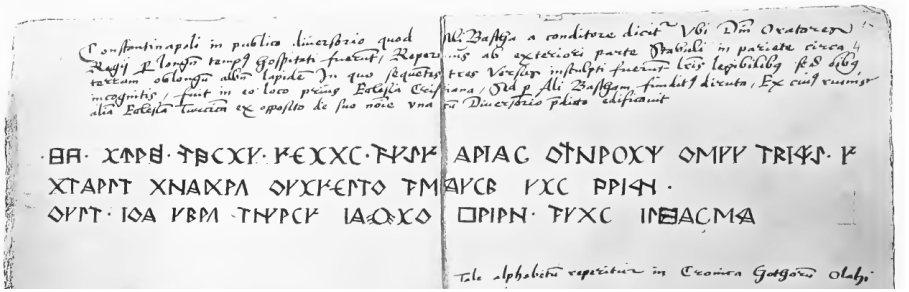


Figure 8: Copy of the original mural inscription in Constantinople (Sebestyén 1915)

	·ⵒⵒ· ⵅⵜⵔⵒ· ⵖⵒⵛⵅⵢ· ⵔⵑⵑ ⵃⵔⵏⵃ ⵔⵏⵏⵔⵅⵅ ⵔⵏⵔⵔ ⵖⵒⵏⵔ· ⵔ
First row	/ᵃzer efs͡fsᵃāz tizᵉnet ᵃstᵃndēbᵃn īrtākᵃ/īrtᵃk ᵃst lāslō kirāʌērt kᵉvᵃtᵃt jārɔttᵊk ⁱtt/ ‘It was written in 1515; delegate of King Vladislaus was sent here.’
	ⵅⵜⵃⵔⵜ ⵅⵏⵃⵅⵔⵃ ⵔⵅⵅⵑⵑⵜⵔ ⵒⵜⵃⵃⵔⵛⵒ ⵒⵅⵛ ⵒⵒⵏⵏ ·
Second row	/bīlɔji bᵊrlɔbāš kᵃt ästᵃndᵃjik it vɔlt; nᵃm tēn f͡ʃāsār/ ‘Barnabas Bélay waited here for two years; the emperor did not do [anything for them].’
	ⵔⵅⵜⵜ· ⵏⵔⵃ ⵅⵒⵒⵃ ⵒⵅⵔⵑ ⵏⵃⵔⵛⵅ ⵏⵃⵃⵔⵅⵔ □ⵒⵏⵑ· ⵒⵅⵛ ⵏⵒⵒⵃⵛⵑⵃ
Third row	/kⁱdᵃji sᵉkᵉl tᵊmāš īrtān ᵃst, sᵉlⁱmbēk f͡ʃāsār idᵃ tēn sāz lōvɔl/ ‘Thomas Kidei Székely wrote here, Emperor Selim housed here with one hundred horses.’

Table 17: The transliteration, transcription, and translation of the Constantinople inscription

by the fact that in the word ⵏⵏ /īrtāk/īrtᵃk/ ‘written’ the /ā/ is not written with an individual grapheme; however, the long vowels were generally written even in the early Rovash inscriptions (Zelliger and Hosszú 2014).

Bibliography

Adiego Lajara, Ignasi-Xavier. [Ignacio-Javier Adiego Lajara.] “Los alfabetos epicóricos anhelénicos de Asia Menor.” In Bádenas de la Peña, Pedro et al. (eds.). *Lenguas en contacto: el testimonio escrito.* Madrid: Consejo Superior de Investigaciones Científicas, 2004. 299–320.

Adiego Lajara, Ignasi-Xavier. [2007a.] *The Carian language.* Leiden: Koninklijke Brill, 2007.

Adiego Lajara, Ignasi-Xavier. [2007b.] “Greek and Carian.” In Christidis, Anastassios-Fivos (ed.). *A History of Ancient Greek. From the Beginnings to Late Antiquity.* Cambridge, New York (NY): Cambridge University Press, 2007. 758–762.

Adiego Lajara, Ignasi-Xavier. [2007c.] "Greek and Lycian." In Christidis, Anastassios-Fivos (ed.). *A History of Ancient Greek. From the Beginnings to Late Antiquity.* Cambridge, New York (NY): Cambridge University Press, 2007. 763–767.

Adiego Lajara, Ignasi-Xavier. [2007d.] "Greek and Lydian." In Christidis, Anastassios-Fivos (ed.). *A History of Ancient Greek. From the Beginnings to Late Antiquity.* Cambridge, New York (NY): Cambridge University Press, 2007. 768–772.

Adiego Lajara, Ignasi-Xavier. [2007e.] *The Spread of Alphabetic Writing among the Non-Greek Peoples of Anatolia.* Unpublished, 2007.

Adiego Lajara, Ignasi-Xavier. "Lycian nasalized preterites revisited." *Indogermanische Forschungen* 120 (2015). 1–30.

d'Alfonso, Lorenzo and Annick Payne. "The Paleography of Anatolian Hieroglyphic Stone Inscriptions." *Journal of Cuneiform Studies* 68 (2016). 107–127.

Anders, Gunter. *Luwhitta/B. Luwische Hieroglyphen Fonts.* 2012. <http://www.hethport.uni-wuerzburg.de/luwglyph>.

Aussems, Johannes F. A. "Christine de Pizan and the scribal fingerprint—A quantitative approach to manuscript studies." Utrecht: unpublished Research Master's Thesis, 2006. <http://dspace.library.uu.nl/handle/1874/12537>.

Aussems, Mark, and Axel Brink. "Digital Palaeography." In Rehbein, Malte, Patrick Sahle, and Torsten Schaßen (eds.). *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age (KPDZ 1).* Norderstedt: Books on Demand, 2009. 293-308.

Babinger, Franz. "Eine neuentdeckte ungarische Kerbinschrift aus Konstantinopel vom Jahre 1515." *Ungarische Rundschau* III (1914). 41–52.

Baeza-Yates, Ricardo A. "Introduction to data structures and algorithms related to information retrieval." In Frakes, William B., and Ricardo A. Baeza-Yates (eds.). *Information Retrieval: Data Structures and Algorithms.* Upper Saddle River (NJ): Prentice-Hall, Inc, 1992. 13-27.

Bakkum, Gabriël C. L. M. *The Latin Dialect of the Ager Faliscus. 150 Years of Scholarship. Part I.* Amsterdam: University of Amsterdam, 2009.

Bálint, Csanád. *Der Schatz von Nagyszentmiklós: archäologische Studien zur frühmittelalterlichen Metallgefäßkunst des Orients, Byzanz' und Steppe.* Budapest: Balassi Kiadó, 2010.

Bar-Joseph, Ziv, David K. Gifford, and Tommi S. Jaakkola. "Fast optimal leaf ordering for hierarchical clustering." *Bioinformatics* 17 (2001). S22–S29.

Barbançon, François et al. "An experimental study comparing linguistic phylogenetic reconstruction methods." *Diachronica* 30 (2013). 143-170.

Benkő, Elek. "Középkori rovásfelirat Vargyasról." [Medieval Rovash inscription of Vargyas.] *Magyar Nyelv* 90 (1994). 487-489.

Benkő Elek. [1996a.] "A székely rovásírás korai emlékei." [The early relics of the Székely Rovash script.] *Magyar Nyelv* 92 (1996). 75–80.

Benkő Elek. [1996b.] "A székely rovásírás. A legújabb kutatások." [The Székely Rovash script. The latest research.] *História* 18 (1996). 31–33.

Bernal, Martin. *Cadmean Letters. The Transmission of the Alphabet to the Aegean and Further West before 1400 B.C.* Winona Lake (IN): Eisenbrauns, 1990.

Berry, Michael J. A. and Gordon Linoff. *Data Mining Techniques for Marketing. Sales and*

Customer Support. New York (NY): John Wiley & Sons, Inc, 1996.
Beyer, Klaus. *Die aramäischen Inschriften aus Assur, Hatra und dem übrigen Ostmesopotamien: (datiert 44. v.Chr. bis 238 n.Chr.).* Göttingen: Vandenhoeck & Ruprecht, 1998.
Boisson, Claude. "Conséquences phonétiques de certaines hypotheses de déchiffrement du carien." In Gianotta, Maria Eliana et al. (eds.). *La decifrazione del Cario.* Rome: Consiglio nazionale delle ricerche, 1994. 207–232.
Bordreuil, Pierre. "Migraines d'Épigraphiste." In Bienkowski, Piotr, Christopher Mee, and Elizabeth Slater (eds.). *Writing and Ancient Near Eastern Society. Papers in Honour of Alan R. Millard.* New York (NY), London: T&T Clark International, 2005. 15–28.
Botto, Massimo. "The Phoenicians in the central-west Mediterranean and Atlantic between 'precolonization' and the 'first colonization'." In Donnellan, Lieve, Valentino Nizzo, and Gert-Jan Burgers (eds.). *Contexts of Early Colonization.* Rome: Palombi Editori, 2016. 289–309.
Brixhe, Claude and Michel Lejeune. *Corpus des inscriptions paléo-phrygiennes.* Paris: Éditions Recherche sur les civilisations, 1984.
Christidis, Anastassios-Fivos (ed.). *A History of Ancient Greek. From the Beginnings to Late Antiquity.* Cambridge (NY): Cambridge University Press, 2007.
Ciula, Arianna. "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis." *Digital Medievalist* 1 (2005). DOI: 10.16995/dm.4.
Ciula, Arianna. "The Palaeographical Method Under the Light of a Digital Approarch." In Rehbein, Malte, Patrick Sahle, and Torsten Schaßen (eds.). *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age (KPDZ 1).* Norderstedt: Books on Demand, 2009. 219–235.
Cloppet, Florence et al. "New Tools for Exploring, Analysing and Categorising Medieval Scripts." *Digital Medievalist* 7 (2011). <http://www.digitalmedievalist.org/journal/7/cloppet>.
Colless, Brian E. "Proto-alphabetic inscriptions from the Wadi Arabah." *Antiguo Oriente* 8 (2010). 75–96.
Correia, Virgílio Hipólito. *A epigrafia da Idade do Ferro do Sudoeste da Península Ibérica.* Porto: Etnos, 1996.
Coulmas, Florian. *The Blackwell Encyclopedia of Writing Systems.* Oxford: Blackwell, 1999.
Cross, Frank M. "The Invention and Development of the Alphabet." In Senner, Wayne M. (ed.). *The Origins of Writing.* Lincoln (NE): University of Nebraska Press, 1989. 77–90.
Daniels, Peter T. and William Bright (eds.). *The World's Writing Systems.* New York (NY), Oxford: Oxford University Press, 1996.
Davies, Anna Morpurgo, and Jean-Pierre Olivier. "Syllabic scripts and languages in the second and first millennia BC." *British School at Athens Studies* 20 (2012). 105–118.
Davis, Brent. "Introduction to the Aegean pre-alphabetic scripts." *Kubaba* 1 (2010). 38–61.
Doblhofer, Ernst. *Jelek és csodák.* [Signs and wonders.] Budapest: Gondolat, 1962.
Dunn, Joe C. "Well separated clusters and optimal fuzzy partitions." *Journal of Cybernetics* 4 (1974). 95–104.
Egetmeyer, Markus. *Le dialecte grec ancien de Chypre.* Berlin; New York (NY): de Gruyter, 2010.
Enoki, Kazuo, Gennadij A. Koshelenko, and Z. Haidary. "The Yüeh-chih and their Migrations." In Harmatta, János (ed.). *History of Civilizations of Central Asia: The development of sedentary*

and nomadic civilizations: 700 B.C. to A.D. 250. Volume II. Paris: UNESCO Publishing, 1994. 165–183.

Erdal, Marcel. "The runic graffiti at Yar Khoto." *Türk Dilleri Araştırmaları* 3 (1993). 87–108.

Erdal, Marcel. *A Grammar of Old Turkic.* Leiden: Koninklijek Brill, 2004.

Erdal, Marcel. *Personal e-mail communication.* 7 Nov. 2016.

Eska, Joseph F. "Continental Celtic." In Woodard, Roger D. (ed.). *The Ancient Languages of Europe.* Cambridge: Cambridge University Press, 2008. 165–188.

Farrujia de la Rosa, A. José et al. "The Libyco-Berber and Latino-Canarian Scripts and the Colonization of the Canary Islands." *African Archaeological Review* 27 (2010). 13–41.

Faulmann, Carl. *Das Buch der Schrift. Enthaltend die Schriftzeichen und Alphabete aller Zeiten und aller Völker des Erdkreises.* Wien: Druck und Verlag der Kaiserlichen Hof- und Staatsdruckerei, 1880.

Ferrer i Jané, Joan. "Novetats sobre el sistema dual de diferenciació gràfica de les oclusives sordes i sonores." *Acta Palaeohispanica* 5 (2005). 957–982.

Ferrer i Jané, Joan. "Els sistemes duals de les escriptures ibèriques." *Palaeohispanica* 13 (2013). 445–459.

Ferrer i Jané, Joan. "Ibèric *kutu* i els abecedaris Ibèrics [Iberian kutu and the Iberian Abecedaries.]" *Veleia* 31 (2014). 227–259.

Fischer, Franz, Christiane Fritze, and Georg Vogeler (eds.). *Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2 (KPDZ 2).* Norderstedt: Books on Demand, 2010.

Fischer, Steven R. *A History of Writing.* London: Reaktion Books, 2001.

Forster, Peter and Colin Renfrew (eds.). *Phylogenetic Methods and the Prehistory of Languages.* Cambridge: McDonald Institute for Archaeological Research, 2006.

Gabain, Annemarie von. *Alttürkische Grammatik mit Bibliographie, Lesestücken und Wörterverzeichnis auch Neutürkisch. Mit vier Schrifttafeln und sieben Schriftproben.* Leipzig: Harrassowitz, 1941.

Garbini, Giovanni. *Storia e problemi dell'epigrafia semitica.* Naples: Istituto Orientale, 1979.

Gibson, John C. L. *Textbook of Syrian Semitic Inscriptions.* Vol. 2. Oxford: Oxford University Press, 1975.

Goodall, David W. "Objective methods for the classification of vegetation III. An essay in the use of factor analysis." *Australian Journal of Botany* 2 (1954). 304–324.

Gower, John C. and Pierre Legendre. "Metric and Euclidean Properties of Dissimilarity Coefficients." *Journal of Classification* 3 (1986). 5–48.

Grimme, Hubert. *Althebräische Inschriften vom Sinai: Alphabet, Textliches, Sprachliches mit Folgerungen.* Hannover: Heinz Lafaire, 1923.

Hampel, József. "A nagyszentmiklósi kincs. Tanulmány a népvándorláskori művészetről." [The treasure of Nagyszentmiklós. A study of the art of the Great Migrations.] *Archeológiai Értesítő* 4 (1884). 1–166.

Harmatta, János. "Conclusion." In Harmatta, János (ed.). *History of Civilizations of Central Asia: The development of sedentary and nomadic civilizations: 700 B.C. to A.D. 250.* Volume II. Paris: UNESCO Publishing, 1994. 476–483.

Hawkins, John David. "Writing in Anatolia: Imported and Indigenous Systems." *World Archae-*

ology 17 (1986). 363–376.
Hawkins, John David. *Corpus of Hieroglyphic Luwian Inscriptions. Vol. I: Inscriptions of the Iron Age.* Berlin; New York (NY): de Gruyter, 2000.
Hawkins, John David. "A Unique Hieroglyphic Luwian Document?" In Cohen, Yoran, Amir Gilan, and Jared L. Miller (eds.). *Pax Hethitica – Studies on the Hittites and their Neighbours in Honour of Itamar Singer.* Wiesbaden: Harrassowitz, 2010. 183–190.
Healey, John F. *The Early Alphabet.* Berkeley, Los Angeles (CA): University of California Press, 1990.
Hempl, George. "The Origin of the Latin Letters G and Z." *Transactions and Proceedings of the American Philological Association* 30 (1899). 24–41.
Hennig, Willi. *Phylogenetic systematics.* Urbana (IL): University of Illinois Press, 1966.
Hesperia: *Banco de datos de lenguas paleohispánicas.* <http://hesperia.ucm.es>.
Hosszú, Gábor and Erzsébet Zelliger. "Rovásfelirat a megfejtés útján. Módszertani kérdések." [On the deciphering of Rovash inscriptions. Methodological questions.] *E-nyelv Magazin* 2 (2013). <http://e-nyelvmagazin.hu/rovasfelirat-a-megfejtes-utjan-modszertani-kerdesek>.
Hosszú, Gábor and Erzsébet Zelliger. [2014a.] "Többnyelvű feliratok a Nagyszentmiklósi aranykincsen." [Multilingual inscriptions on the golden treasure of Nagyszentmiklós.] *Magyar Nyelv* 110 (2014). 177–195. <http://www.c3.hu/~magyarnyelv/14-2/HosszuZelliger14-2.pdf>.
Hosszú, Gábor and Erzsébet Zelliger. [2014b.] "A Bodrog-alsóbűi rovásemlék számítógépes írástörténeti kapcsolatai és egy olvasati kísérlete." [Computational palaeographical relations of the Bodrog-Alsóbű Rovash relic and a reading attempt.] *Magyar Nyelv* 110 (2014). 417–431. <http://www.c3.hu/~magyarnyelv/14-4/HG_ZE_2014-4.pdf>.
Hosszú, Gábor. *Heritage of Scribes. The Relation of Rovas Scripts to Eurasian Writing Systems.* Budapest: Rovas Foundation, 2012.
Hosszú, Gábor. *Rovásatlasz.* [Rovash Atlas.] Budapest: Milani, 2013.
Hosszú, Gábor. "Mathematical Statistical Examinations on Script Relics." In Bhatnagar, Vishal (ed.). *Data Mining and Analysis in the Engineering Field.* Hershey (PA): Information Science Reference, 2014. 142–158.
Hosszú, Gábor. "A Novel Computerized Paleographical Method for Determining the Evolution of Graphemes." In Khosrow-Pour, Mehdi (ed.). *Encyclopedia of Information Science and Technology.* Hershey (PA): Information Science Reference, 2015. 2017–2031.
Jain, Anil K. and Richard C. Dubes. *Algorithms for Clustering Data.* Upper Saddle River (NJ): Prentice-Hall, 1988.
Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data Clustering: A Review." *ACM Computing Surveys* 31 (1999). 264-323.
Jeffery, Lilian H. *The local scripts of archaic Greece.* Oxford: Clarendon Press, 1961.
Jensen, Hans. *Sign, Symbol, and Script: An Account of Man's Efforts to Write.* Transl. George Unwin. New York (NY): Putman, 1969.
Kairzhanov, Abai K. [Каиржанов, Абай Каиржанович.] Генезис тюркского рунического письма (на материале тамги-знаков евразийских кочевников). *Мова* 22 (2014). 12–24.
Kalinka, Ernst. *Tituli Asiae Minoris.* Vol. 1: *Tituli Lyciae lingua Lycia conscripti.* Wien: Hölder,

1901.

Kara, György. "Aramaic Scripts for Altaic Languages." In Daniels, Peter T., and William Bright (eds.). *The World's Writing Systems.* New York (NY), Oxford: Oxford University Press, 1996. 536–558.

Karali, Maria. "Writing systems." In Christidis, Anastassios-Fivos (ed.). *A History of Ancient Greek. From the Beginning to Late Antiquity.* Cambridge (NY): Cambridge University Press, 2007. 197–207.

Karnava, Artemis. *The Cretan Hieroglyphic script of the second millennium BC: description, analysis, function and decipherment perspectives.* Unpublished doctoral dissertation. Bruxelles: Université libre de Bruxelles, 1999.

Kaufman, Leonard and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* New York (NY): John Wiley & Sons, 1990.

Kenyon, Frederic G. *The Palaeography of Greek Papyri.* Oxford: Clarendon Press, 1899.

King, Geraldine. *A Report on the Work of the Dhofar Epigraphic Project.* <http://krc2.orient.ox.ac.uk/aalc/images/documents/mcam/dep_report_1992.pdf>.

Kitching, Ian J. et al. *Cladistics: The Theory and Practice of Parsimony Analysis.* Oxford: Oxford University Press, 1998.

Kloekhorst, Alwin. "Studies in Lycian and Carian phonology and morphology." *Kadmos* 47 (2008). 117–146.

Konkobaev, Kadyraly, Nurdin Useev, and Negizbek Shabdanaliev [Кадыралы Конкобаев, Нурдин Усеев, Негизбек Шабданалиев]. *Атлас древнетюркских письменных памятников Республики Алтай.* Astana: Ғылым, 2015.

Kononov, Andrej Nikolaevič. *Grammatika jazyka tjurkskix runičeskix pamjatnikov.* Leningrad: Nauka, 1980.

Krings, Véronique (ed.). *La civilisation phénicienne et punique: manuel de recherche.* Leiden, New York (NY), Köln: E.J. Brill, 1995.

Kyzlasov, Igor L. (Кызласов, Игорь Леонидович). *Руническая письменности евразийских степей*, Москва: Издательская фирма «Восточная литература» РАН, 1994.

LBI: *Libyco-Berber Inscriptions Online Database. © LBI-Projekt / Institutum Canarium* <http://www.institutum-canarium.org>.

Lemaire, André and Benjamin Sass. "The Mortuary Stele with Sam'alian Inscription from Ördekburnu near Zincirli." *BASOR* 369 (2013). 57–136.

Lidzbarski, Mark. "Der Ursprung der nord-und südsemitischen Schrift." In Lidzbarski, Mark (ed.). *Ephemeris für semitische Epigraphik. Vol. I. 1900–1902.* Giessen: Ricker, 1902. 109–136.

Ligeti, Lajos. "A magyar rovásírás egy ismeretlen betűje." [An unknown letter of the Hungarian Rovash script.] *Magyar Nyelv* XXI (1925). 50–52.

Lindberg, David R. *Principals of Phylogenetic Systematics: Phenetics. Integrative Biology 200A Principles of Phylogenetics: Systematics.* University of Berkeley, 2012. <http://ib.berkeley.edu/courses/ib200a/lect/ib200a_lect09_Lindberg_phenetics.pdf>.

Looijenga, Tineke. *Runes Around the North Sea and On the Continent AD 150–700; Texts & Contexts.* Groningen: SSG Uitgeverij, 1997.

Macdonald, Michael C. A. "Reflections on the linguistic map of pre-Islamic Arabia." *Arabian*

archaeology and epigraphy 11 (2000). 28–79.
Macdonald, Michael C. A. "Ancient North Arabian." In Woodard, Roger D. (ed.). *The Cambridge Encyclopedia of the World's Ancient Languages.* Cambridge: Cambridge University Press, 2004. 488–533.
Macdonald, Michael C. A. "Literacy in an Oral Environment." In Bienkowski, Piotr, Christopher Mee, and Elizabeth Slater (eds.). *Writing and Ancient Near Eastern Society. Papers in Honour of Alan R. Millard.* New York (NY), London: T&T Clark International, 2005. 45–114.
Macdonald, Michael C. A. *Literacy and Identity in Pre-Islamic Arabia.* Farnham, Burlington: Ashgate Publishing, 2009.
Macdonald, Michael C. A. "Ancient Arabia and the written word." In Macdonald, Michael C.A. (ed.). *The development of Arabic as a written language.* (Supplement to the Proceedings of the Seminar for Arabian Studies 40). Oxford: Archaeopress, 2010. 5–28.
Macdonald, Michael C. A. "On the uses of writing in ancient Arabia and the role of palaeography in studying them." *Arabian Epigraphic Notes* 1 (2015). 1–50.
MacKenzie, David Neil. *A Concise Pahlavi Dictionary.* London: Oxford University Press, 1971.
Marchesini, Simona. *Le lingue frammentarie dell'Italia antica: manuale per lo studio delle preromane.* Milano: Hoepli, 2009.
Marchesini, Simona. "The Elymian language." In Tribulato, Olga (ed.). *Language and Linguistic Contact in Ancient Sicily.* Cambridge: Cambridge University Press, 2012. 95–114.
Marchesini, Simona. "Über die Rätische Inschrift aus Pfatten/Vadena im Tiroler Landesmuseum Ferdinandeum, Innsbruck." In Meighörner, Wolfgang (ed.). *Wissenschaftliches Jahrbuch der Tiroler Landesmuseen 2014.* Innsbruck, Wien, Bozen: StudienVerlag, 2014. 202–217.
Marsadolov, Leonid. [2000a.] "The Cimmerian Traditions of the Gordion Tumuli (Phrygia): Found in the Altai Barrows (Bashadar, Pazyryk)." In Davis-Kimball, Jeannine et al. (eds.). *Kurgans, Ritual Sites, and Settlements: Eurasian Bronze and Iron Age.* Oxford: Archaeopress, 2000. 247–258.
Marsadolov, Leonid. [2000b.] "The Nomads of Kazakhstan, the Altai, and Tuva." In Aruz, Joan et al. (eds.). *The Golden Deer of Eurasia. Scythian and Sarmatian Treasures from the Russian Steppes.* New York (NY): Yale University Press, 2000. 49–53.
Masson, Olivier. "Un lion de bronze de provenance égyptienne avec inscription carienne." *Kadmos* 15 (1976). 82–83.
Masson, Olivier. *Carian Inscriptions from North Saqqâra and Buhen.* London: Egypt Exploration Society, 1978.
MATLAB: *MATLAB Release R2015a.* Natick (MA): The MathWorks, Inc., 2015.
Maue, Dieter. "Three Languages on One Leaf: On IOL Toch 81 with Special Regard to the Turkic Part." *Bulletin of the School of Oriental and African Studies, University of London* 71 (2008). 59–73.
McCarter, P. Kyle, Jr. *The antiquity of the Greek alphabet and the early Phoenician scripts.* Missoula (MT): Scholars Press, 1975.
Mees, Bernard. "Runes in the First Century." In Stoklund, Marie et al. (eds.). *Runes and their Secrets. Studies in runology.* Copenhagen: Museum Tusculanum Press, 2006. 201–232.
Melchert, H. Craig. "Lydian." In Woodard, Roger D. (ed.). *The Cambridge Encyclopedia of the World's Ancient Languages.* Cambridge: Cambridge University Press, 2004. 601–607.

Melchert, H. Craig. [2008a.] "Lycian." In Woodard, Roger D. (ed.). *The Ancient Languages of Asia Minor.* Cambridge: Cambridge University Press, 2008. 46–55.

Melchert, H. Craig. [2008b.] "Lydian." In Woodard, Roger D. (ed.). *The Ancient Languages of Asia Minor.* Cambridge: Cambridge University Press, 2008. 56–63.

Melchert, H. Craig. [2008c.] "Carian." In Woodard, Roger D. (ed.). *The Ancient Languages of Asia Minor.* Cambridge: Cambridge University Press, 2008. 64–68.

Miller, D. Gary. *Ancient Scripts and Phonologial Knowledge.* Amsterdam, Philadelphia (PA): John Benjamins Publishing, 1994.

MNAMON: *Antiche Scritture del Mediterraneo. Guida critica alle risorse elettroniche.* <http://lila.sns.it/mnamon>.

Morandi, Alessandro. *Epigrafia italica.* Roma: L'Erma di Bretschneider, 1982.

Morandi, Alessandro. "Epigrafia e lingua dei Celti d'Italia." In Agostinetti, Paola P., and Alessandro Morandi (eds.). *Celti d'Italia.* Vol. II. Roma: Spazio Tre, 2004.

Nagy, Géza. "A székely írás eredete." [The origin of the Székely script.] *Ethnographia* VI (1895). 269–276.

Nakhleh, Luay et al. "A comparison of phylogenetic reconstruction methods on an IE dataset." *Transactions of the Philological Society* 3 (2005). 171–192.

Németh, Gyula. "A régi magyar írás eredete." [The origin of the Ancient Hungarian script.] *Nyelvtudományi Közlemények* XLV (1917–1920). 31–44.

Németh, Gyula. *A magyar rovásírás.* [The Hungarian Rovash script.] Budapest: Hungarian Academy of Sciences, 1934.

NLR: *Российская национальная библиотека.* [National Library of Russia.] <http://expositions.nlr.ru/slav_culture>.

Nollé, Johannes. *Side im Altertum. Geschichte und Zeugnisse.* Band II. Bonn: Dr. Rudolf Habelt, 2001.

O'Connor, Michael P. "Epigraphic Semitic Scripts." In Daniels, Peter T. and William Bright (eds.). *The World's Writing Systems.* New York (NY), Oxford: Oxford University Press, 1996. 88–107.

Olivier, Jean-Pierre and Frieda Vandenabeele. *Édition holistique des textes chypro-minoens (HoChyMin).* Pisa; Rome: Fabrizio Serra Editore, 2007.

Olivier, Jean-Pierre. "Les syllabaires chypriotes des deuxième et premier millénaires avant notre ère: État des questions." In Sacconi, Anna et al. (eds.). *Colloguium Romanum. Atti del XII colloquio internazionale di Micenologia, Roma 20–25 febbraio 2006.* Pisa; Rome: Fabrizio Serra, 2008. 605–619.

Pardede, Raymond E. I. et al. "Four-Layer Grapheme Model for Computational Paleography." *Journal of Information Technology Research* 9 (2016). 64–82.

Payne, Annick. *Hieroglyphic Luwian. An Introduction with Original Texts.* Wiesbaden: Harrasowitz, 2010.

Podani, János. *Introduction to the Exploration of Multivariate Biological Data.* Leiden: Backhuys Publishers, 2000.

Praetorius, Franz. "Bemerkungen zum südsemitischen Alphabet." *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 58 (1904). 715–726.

PROEL: *Promotora Española de Lingüística.* <http://www.proel.org>.

Rehbein, Malte, Patrick Sahle, and Torsten Schaßan (eds.) *Kodikologie und Paläographie im digitalen Zeitalter 1 – Codicology and Palaeography in the Digital Age 1 (KPDZ 1).* Norderstedt: Books on Demand, 2009. URN: urn:nbn:de:hbz:38-29393.

Rilly, Claude and Alex de Voogt. *The Meroitic Language and Writing System.* Cambridge: Cambridge University Press, 2012.

Rodríguez Ramos, Jesús. "El origen de la escritura sudlusitano-tartesia y la formación de alfabetos a partir de alefatos." *Rivista di Studi Fenici.* 30 (2002). 187–216.

Rodríguez Ramos, Jesús. *Análisis de epigrafía íbera.* Vitoria-Gasteiz: Universidad del País Vasco, 2004.

Rogers, Henry. "Sociolinguistic factors in borrowed writing systems." *Toronto Working Papers in Linguistics* 17 (1999). 247–262.

Rollston, Christopher A. "The Phoenician Script of the Tel Zayit Abecedary and Putative Evidence for Israelite Literacy." In Tappy, Ron E. and P. Kyle McCarter Jr. (eds.). *Literate Culture and Tenth-Century Canaan. The Tel Zayit Abecedary in Context.* Winona Lake (IN): Eisenbrauns, 2008. 61–96.

Romesburg, H. Charles. *Cluster Analysis for Researchers.* Raleigh: Lulu.com, 2004.

Róna-Tas, András. "On the development and origin of the East Turkic «runic» script." *Acta Orientalia Academiae Scientiarum Hungaricae* XLI (1987). 7–14.

Rosenthal, Franz, Jonas C. Greenfield, and Shaul Shaked. "Aramaic." *Encyclopædia Iranica* II (3) (1986, updated in 2011). 250–261. <http://www.iranicaonline.org/articles/aramaic-#pt2>.

Röllig, Wolfgang. "L'Alphabet." In Krings, Véronique (ed.). *La civilisation phénicienne et punique: Manuel de recherche.* Leiden, New York (NY), Köln: E.J. Brill, 1995. 193–214.

Saitou, Naruya, and Masatoshi Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* 4 (1987). 406–425.

Sampaio, Alberto. "Software Phenetics, Phylogeny and Evolution." *Proceedings of the 3rd International IEEE Workshop on Software Evolvability* 2007. 60–66.

Sándor, Klára. *A Bolognai Rovásemlék.* [The Bologna Rovash relic.] Szeged: József Attila Tudományegyetem, Magyar Őstörténeti Kutatócsoport, 1991.

Sass, Benjamin. *The Genesis of the Alphabet and Its Development in the Second Millennium B.C.* Wiesbaden: Harrassowitz, 1988.

Sebestyén, Gyula. *Rovás és rovásírás* [Rovash and Rovash writing.] Budapest: Magyar Néprajzi Társaság, 1909.

Sebestyén, Gyula. *A magyar rovásírás hiteles emlékei.* [The credible relics of the Hungarian Rovash script.] Budapest: Hungarian Academy of Sciences, 1915.

Segert, Stanislav. "Vowel Letters in Early Aramaic." *Journal of Near Eastern Studies* 37 (1978). 111–114.

Simon, Zsolt. "Critica. Ignacio J. Adiego: The Carian Language. With an Appendix by Koray Konuk [Handbuch der Orientalistik 86]. Leiden; Boston (MA): Brill, 2007. xiv + 518 pages with 2 maps & 4 plates." *Acta Antiqua Academiae Scientiarum Hungaricae* 48 (2008). 457–463.

Sims-Williams, Nicholas. "The Sogdian Sound-System and the Origins of the Uyghur Script." *Journal Asiatique* (1981). 347–360.

Sims-Williams, Nicholas. "Sogdian." In Schmitt, Rüdiger (ed.). *Compendium Linguarum Irani-*

carum. Wiesbaden: Dr. Ludwig Reichert Verlag, 1989. 173-192.
Sims-Williams, Nicholas. "The Sogdian manuscripts in Brāhmī script as evidence for Sogdian phonology." In Emmerick, Ronald E. et al. (eds.). *Turfan, Khotan und Dunhuang. Vorträge der Tagung Annemarie von Gabain und die Turfanforschung, Berlin, 9.–12. 12. 1994.* Berlin: Akademie Verlag, 1996. 307–315.
Sims-Williams, Nicholas and Franz Grenet. "The Sogdian Inscriptions of Kultobe." *Shygys* 2006/1 (2007). 95–111.
Skelton, Christina. "Phylogenetic Systematics to Reconstruct the History of the Linear B Script." *Archaeometry* 50 (2007). 158–176.
Skjærvø, Prods Oktor. "Aramaic Scripts for Iranian Languages." In Daniels, Peter T. and William Bright (eds.). *The World's Writing Systems.* New York (NY), Oxford: Oxford University Press, 1996. 515–535.
Sneath, Peter H. A. and Robert Reuven Sokal. *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* San Francisco (CA): Freeman, 1973.
Sneath, Peter H. A. "Cladistic Representation of Reticulate Evolution." *Systematic Zoology* 24 (1975). 360–368.
Sokal, Robert Reuven and Charles D. Michener. "A statistical method for evaluating systematic relationships." *The University of Kansas Science Bulletin* XXXVIII (1958). 1409–1438.
Sokal Robert Reuven and Peter H. A. Sneath. *Principles of Numerical Taxonomy.* San Francisco (CA), London: Freeman, 1963.
Sokal, Robert Reuven et al. "Numerical Taxonomy: Some Points of View." *Systematic Zoology* 14 (1965). 237–243.
Sprengling, Martin. *The Alphabet. Its Rise and Development from the Sinai Inscriptions.* Chicago (IL): University of Chicago Press, 1931.
Stokes, Peter Anthony. "Computer-Aided Palaeography, Present and Future." In Rehbein, Malte, Patrick Sahle, and Torsten Schaßen (eds.). *Kodikologie und Paläographie im digitalen Zeitalter 1 – Codicology and Palaeography in the Digital Age 1 (KPDZ 1).* Norderstedt: Books on Demand, 2009. 309–338.
Stokes, Peter Anthony. "Digital Approaches to Paleography and Book History: Some Challenges, Present and Future." *Frontiers in Digital Humanities* 2 (2015). 1–3.
Sukkarieh, Jana Z., Matthias von Davier, and Kentaro Yamamoto. *From Biology to Education: Scoring and Clustering Multilingual Text Sequences and Other Sequential Tasks.* Princeton (NJ): Educational Testing Service, ETS Research Report No. RR-12–25, Dec. 2012.
Swiggers, Pierre. "Transmission of the Phoenician Script to the West." In Daniels, Peter T. and William Bright (eds.). *The World's Writing Systems.* New York (NY), Oxford: Oxford University Press, 1996. 261–270.
Swiggers, Pierre and Wolfgang Jenniges. "The Anatolian Alphabets." In Daniels, Peter T. and William Bright (eds.). *The World's Writing Systems.* New York (NY); Oxford: Oxford University Press, 1996. 281–287.
Szabó, Géza. "Keleti mítoszelemek nyugati megjelenése a regölyi astralagos leletek tükrében." [Occurance of eastern mythological elements in the West considering the astralag finds of Regöly.] In Csabai, Zoltán et al. (eds.). *Európé égisze alatt. Ünnepi tanulmányok Fekete Mária hatvanötödik születésnapjára kollégáitól, barátaitól és tanítványaitól.* [Under the

perspective of Európé. Celebrating studies for the 65th birthday of Mária Fekete from her collegues, friends and disciples.] Budapest: L'Harmattan, 2015. 321–364.

Taylor, Isaac. *The Alphabet. An Account of the Origin and Development of Letters. Vol. II. Aryan Alphabets.* London: Kegan Paul, Trench, & Co., 1883.

Tekin, Talat. *Orhon Türkçesi Grameri.* Istanbul: Mehmet Ölmez, 2003.

Thelegdi, Ioannis. *Rudimenta, Priscae hunnorum linguae brevibus quaestionibus ac responcionibus comprehensa opera et studio. (1598).* Budapest: Ars Libri, 1994.

Thompson, Edward M. *An Introduction to Greek and Latin Palaeography.* Oxford: Clarendon Press, 1912.

Thomsen, Vilhelm. *Inscriptions de l'Orkhon déchiffrées.* Helsingfors: La société de literature Finnoise, 1893.

Tóth, Loránd L., Raymond Pardede, and Gábor Hosszú. "Novel Algorithmic Approach to Deciphering Rovash Inscriptions." In Khosrow-Pour, Mehdi (ed.). *Encyclopedia of Information Science and Technology.* Hershey (PA): Information Science Reference, 2015. 7222–7233.

Tóth, Loránd L. et al. "Application of the Cluster Analysis in Computational Paleography." In Samui, Pijush (ed.). *Handbook of Research on Advanced Computational Techniques for Simulation-Based Engineering.* Hershey (PA): Engineering Science Reference, 2016. 525–543.

Tzanavari, Katerina and Anastasios-Phoebus Christidis. "A Carian Graffito from the Lebet Table, Thessaloniki." *Kadmos* 34 (1995). 13–17.

Urbanová, Daniela. "Oština, umberština a jihopikénština – sabellské jazyky." *Sborník prací Filozofické fakulty brněnské univerzity. N, Řada klasická* 52 (2003). 5–37.

Valério, Miguel. "Origin and development of the Paleohispanic scripts: the orthography and phonology of the Southwestern alphabet." *Revista Portuguesa de Arqueologia* 11 (2008). 107–138.

Valério, Miguel. "Cypro-Minoan Tablet RASH Atab 004 as Akkadian Text and its Role in the Decipherment of the Script." In Martín, Sergio C. et al. (eds.). *Mediterráneos: An Interdisciplinary Approach to the Cultures of the Mediterranean Sea.* Newcastle upon Tyne: Cambridge Scholars Publishing, 2013. 5–28.

Valério, Miguel. "«Cypro-Greek» Syllabary: An Unambiguous Solution to a Terminological Problem." *AION Linguistica* 3 (2014). 259–270.

Valério, Miguel. *Investigating the Signs and Sounds of Cypro-Minoan.* Doctoral dissertation, Barcelona: University of Barcelona, 2016.

Vasil'ev, Dmitrij D. "Versuch zur Lösung der Kerbinschrift aus der Umgebund von Kalocsa im Spiegel der eurasischen Parallelen." *Folia Archaeologica* XLIII (1994). 181–191.

Vasil'ev, Dmitrij D. "The Eurasian Areal Aspect of Old Turkic Written Culture." *Acta Orientalia Academiae Scientiarum Hungaricae* 58 (2005). 323–330.

Vékony, Gábor. "Das nordwestliche Transdanubien im 9. Jahrhundert und die 'Uungariorum marcha'." *Savaria* 15 (1981). 215–229.

Vékony, Gábor. [1985a.] "Késő népvándorláskori rovásfeliratok." [Rovash inscriptions from the Late Period of the Great Migrations.] *Életünk* XXII (1985). 71–84.

Vékony, Gábor. [1985b.] "A szarvasi felirat és ami körülötte van." [The Szarvas inscription and its circumstances.] *Életünk* XXII (1985). 1133–1145.

Vékony, Gábor. [1987a.] *Későnépvándorláskori rovásfeliratok a Kárpát-medencében.* [Rovash inscriptions in the Carpathian Basin from the Late Period of the Great Migrations.] Szombathely: Életünk Szerkesztősége – Magyar Írók Szövetsége Nyugat-Magyarországi Csoportja, 1987.

Vékony, Gábor. [1987b.] "Spätvölkerwanderungszeitliche Kerbinschriften im Karpatenbecken." *Acta Archaeologica Academiae Scientiarum Hungaricae* 39 (1987). 211–256.

Vékony, Gábor. [1992a.] "A halomi honfoglalás kori tegezfelirat." [Quiver inscription of Halom from the age of the Magyars' Land Acquisition.] *Életünk* XXX (1992). 542–546.

Vékony, Gábor. [1992b.] "Varázsszöveg a halomi honfoglalás kori temetőből." [Magic text from the cemetery of Halom from the age of the Magyars' Land Acquisition.] In Sándor, Klára (ed.). *Rovásírás a Kárpát-medencében.* [Rovash scripting in the Carpathian Basin.] Szeged: József Attila University, 1992. 41–50.

Vékony, Gábor [1992c.] "A Bolognai Rovásemlék." [The Rovash relic of Bologna.] *Magyar Könyvszemle* 108 (1992). 288–289.

Vékony, Gábor. "A székely rovásírás." [The Székely Rovash script.] *Napjaink* VII (1993). 39.

Vékony, Gábor. "The Peoples of the Period of the Great Migrations in the Carpathian Basin." *Specimina Nova* 12 (1996). 01–04.

Vékony, Gábor. [1999a.] "A székely írás legrégibb emléke Bodrog–Alsóbű vaskohászati műhelyéből." [The oldest relic of the Székely script from Bodrog-Alsóbű iron smelting yard.] In Gömöri, János (ed.). *Traditions and Innovations in the Early Medieval Iron Production*, Sopron – Somogyfajsz, 30th May 1997 – 1st June 1997. Dunaújváros, Veszprém, 1999. 226–229.

Vékony, Gábor. [1999b.] "10. századi székely felirat a Somogy megyei Bodrog határában." [Székely inscription from the 10th c. in the outskirts of Bodrog of Somogy County.] *História* 8 (1999). 30–31.

Vékony, Gábor. *A székely írás emlékei, kapcsolatai, története.* [Relics, relationships and history of the Székely script] Budapest: Nap Kiadó, 2004.

Wallace, Rex E. *The Sabellic Languages of Ancient Italy.* Muenchen: LINCOM, 2007.

Ward, Joe H. Jr. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58 (1963). 236–244.

Weeden, Mark. "Anatolian Hieroglyphs: Logogram vs. Ideogram." In Grodin, Shai (ed.). *Visualizing Knowledge and Creating Meaning in Ancient Writing Systems.* Berlin: PeWe-Verlag, 2014. 81–100.

Wheeler, Ward C., and Peter M. Whiteley. "Historical linguistics as a sequence optimization problem: the evolution and biogeography of Uto-Aztecan languages." *Cladistics* (2014). 1–13.

Wills, Matthew A. "Morphological disparity: A primer." In Adrain, Jonatahn M., Gregory D. Edgecombe, and Bruce S. Lieberman (eds.). *Fossils, Phylogeny, and Form: An Analytical Approach.* New York (NY): Plenum, 2001. 55–144.

Wolf, Lior et al. "Computerized Paleography: Tools for Historical Manuscripts." *18th IEEE International Conference on Image Processing (ICIP).* Brussels, 2011. 3545–3548.

Woodard, Roger D. *Greek Writing from Knossos to Homer.* New York (NY), Oxford: Oxford University Press, 1997.

Woodard, Roger D. *The Textualization of the Greek Alphabet.* New York (NY): Cambridge University Press, 2014.

Woudhuizen, Fred C. "Etruscan Origins: The Epigraphic Evidence." *Talanta* XIV–XV (1982–1983). 91–117.

Woudhuizen, Fred C. [1984–1985a.] "Lydian: Separated from Luwian by Three Signs." *Talanta* XVI–XVII (1984–1985). 91–113.

Woudhuizen, Fred C. [1984–1985b.] "Origins of the Sidetic script." *Talanta* XVI–XVII (1984–1985). 115–126.

Yakubovich, Ilya S. *Sociolinguistics of the Luvian Language.* PhD. Dissertation. Chicago (IL): University of Chicago Press, 2008.

Yakubovich, Ilya S. [2015a.] "The Luwian Language." *Oxford Handbooks Online* (21 Oct. 2015). DOI: 10.1093/oxfordhb/9780199935345.013.18.

Yakubovich, Ilya S. [2015b.] "Phoenician and Luwian in Early Iron Age Cilicia." *Anatolian Studies* 65 (2015). 35–53.

Young, Rodney S. "Old Phrygian Inscriptions from Gordion: Toward a History of the Phrygian Alphabet. (Plates 67–74)." *Hesperia: The Journal of the American School of Classical Studies at Athens* 38 (1969). 252–296.

Younger, John. "Linear A Texts in phonetic transcription & Commentary." (30 November 2000, last update: 26 October 2016). <http://people.ku.edu/~jyounger/LinearA/>.

Younger, John. "Cretan Hieroglyphic Grids: (very tentative suggestions)." (29 July 2003, last update: 21 April 2012). <http://people.ku.edu/~jyounger/Hiero/Hgrids.html>.

Zelliger, Erzsébet. "Nyelvemlék? Művelődéstörténeti emlék? Gondolatok az ÓMS és a vargyasi rovásírásos felirat kapcsán." [Language relic? Cultural historical relic? Thoughts on the Old Hungarian Lament of Mary and the Vargyas Rovash inscription.] In Juhász, Dezső (ed.). *Kerekasztal körül. Huszonöt nyelvészeti tanulmány.* Budapest: Eötvös Loránd University, 2016. 92–98.

Zelliger, Erzsébet, and Gábor Hosszú. "A Konstantinápolyi székely-magyar rovásfelirat számítógépes paleográfiai elemzése." [The computational palaeographical analysis of the Constantinople Székely-Hungarian Rovash inscription.] *Alkalmazott Nyelvtudomány* XIV (2014). 89–124. <https://www.academia.edu/11537828/>.

Prolegomena zu einer digitalen Paläographie des Hieratischen

Svenja A. Gülden, Celia Krause, Ursula Verhoeven

Zusammenfassung

Der folgende Beitrag stammt aus dem Bereich der Ägyptologie und greift ein Thema auf, das Stephen Quirke im zweiten Band der vorliegenden Publikationsreihe erstmals präsentierte. Es geht um die Frage, wie altägyptische Kursivschriften mit digitalen Methoden und unter Berücksichtigung ihres archäologischen und kulturellen Kontexts in Zukunft besser erforscht werden können. Seit 2015 widmet sich das Mainzer Akademievorhaben *Altägyptische Kursivschriften* diesem Komplex und plant eine digitale Paläographie und systematische Analyse des Hieratischen und der Kursivhieroglyphen. Aus den bisherigen Erfahrungen heraus werden im folgenden Beitrag Gegenstand, Methoden, Fragestellungen sowie Kooperationsmöglichkeiten präsentiert und zur interdisziplinären Diskussion gestellt. Enthalten sind grundlegende theoretische Überlegungen, die bei der Erstellung einer digitalen Paläographie für altägyptische kursive Handschriften eine Rolle spielen. Dabei konnten methodische Anregungen aus anderen Disziplinen einbezogen werden. Am Anfang steht eine Einführung in Gegenstand und Methodik der Hieratistik. Danach werden Wege aufgezeigt, wie paläographische Fragen an das Material ins digitale Medium übertragen werden können.

Abstract

The following paper derives from the field of Egyptology and takes up a topic that Stephen Quirke first presented in the second volume of the present series: the question of how ancient Egyptian cursive scripts can be better researched with the help of digital methods in consideration of their archaeological and cultural contexts. Since 2015, a long-term project of the Academy of Literature and Sciences in Mainz under the title of *Altägyptische Kursivschriften* is dedicated to this question. The project develops a digital palaeography on the basis of systematic analysis of the hieratic and cursive hieroglyphic scripts. From experiences gained thus far, methods, problems and questions as well as opportunities for cooperation are now presented and put up for interdisciplinary discussion. Theoretical considerations that are important when composing a digital palaeography for the ancient Egyptian cursive scripts as well as methods used in other disciplines are included. The paper first offers an

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 253–273.

> introduction into the Egyptological subject-matter and the tools and methods used, after which possibilities are proposed on how palaeographical questions and issues can be transferred to the digital medium.[1]

Im Alten Ägypten gab es neben den zumeist gemeißelten Hieroglyphen, die insbesondere für Inschriften auf Monumenten aller Art Verwendung fanden (Abb. 1a), auch kursive (Hand-)Schriften. Diese Kursivschriften, zu denen das Hieratische (Abb. 1b), die Kursivhieroglyphen, das Abnorm- bzw. Kursivhieratische und ab etwa der Mitte des 1. Jahrtausends v. Chr. das Demotische gehören, waren zusammengenommen über 3000 Jahre lang bis in die Römerzeit in Verwendung (von ca. 2800 v. Chr. bis ins 5. Jahrhundert n. Chr.). Man schrieb mit Pflanzenstengeln und Rußtusche (in der Römerzeit mit dem Rohr) überwiegend auf Papyrus, Leinen, Leder, Holz, Ton oder Stein. Die Kursivschriften wurden als erste (und oft wohl auch einzige) Schriftart gelernt und spielten eine wesentliche Rolle für Kommunikation und Verwaltung, für lehrhafte, narrative, fiktionale und poetische Literatur, für Wissensgebiete wie Heilkunde, Mathematik, Astronomie u. a. m. sowie für das weite Feld der religiösen und funerären Texte (Abb. 2; vgl. z. B. Assmann 1994; Parkinson und Quirke 1995; Leach und Tait 2000; Verhoeven 2015b).

Die altägyptischen Kursivschriften benutzen vereinfachte Formen der bildhaften und oft sehr detaillierten Hieroglyphen. Das Hieratische kommt mit etwa 500 bis 600 Hieratogrammen (zum Begriff Verhoeven 2001, 1) – einzelnen Laut- und Deutzeichen, Zahlen, Maßen etc. – aus, während in der Hieroglyphenschrift 700 bis 1000 verschiedene Zeichen (später mit zahlreichen Varianten) verwendet werden. Gerne werden auch zwei oder mehr Zeichen in einer so genannten Ligatur miteinander verbunden. Da als Vorlagen für hieroglyphische Inschriften in der Regel hieratische Texte dienten, fließen mitunter eigenständige hieratische Zeichen, aber auch Verwechslungen aufgrund von Ähnlichkeiten in der hieratischen Schrift in die monumentalen Hieroglyphentexte ein.

1 Paläographien in der Ägyptologie

Auf die Forschungsgeschichte der ägyptologischen Paläographie-Forschung kann im Folgenden nur auszugsweise eingegangen werden (dazu bereits Verhoeven 2015b und Gülden et al., im Druck). Das Standardwerk mit dem Titel *Hieratische Paläographie* wurde zu Beginn des 20. Jahrhunderts von Georg Möller erstellt (Möller 1909-1912), der aus 32 gut datierten Textzeugen, zumeist Papyri, alle unterscheidbaren Einzelzeichen faksimilierte und in übersichtlichen Listen erfasste, in denen jeweils der Bezug zur entsprechenden hieroglyphischen Form gegeben wurde. Die einzelnen Spalten

[1] Wir danken Kyra van der Moezel für die Übersetzung und für verschiedene Diskussionsbeiträge.

Abbildung 1: Zeitgleiche Texte in Hieroglyphen und Hieratisch, ca. 1925 v. Chr. (1a: Weiße Kapelle Sesostris' I., Karnak; 1b: Brief aus dem Hekanachte-Archiv, Papyrus New York MMA 22.3.516)

Abbildung 2: Hieratisch geschriebenes Totenbuch mit Wechsel von roter und schwarzer Tusche sowie bildlichen Darstellungen (Vignetten), ca. 664 – 525 v. Chr. (Papyrus London BM EA 10558.10)

gaben die Textzeugen – chronologisch von links nach rechts fortschreitend – an, während die Zeilen die Beispiele für die kursiven Formen eines Zeichens enthielten. Die Reihenfolge der Zeichen entsprach einer älteren hieroglyphischen Zeichenliste. Erst Alan H. Gardiner entwickelte in seiner *Egyptian Grammar* (Gardiner 1927) die heutige Standardliste der Ägyptologie. Möllers drei Bände mit zusammengenommen etwa 220 Seiten deckten die Zeitspanne von der 5. Dynastie (ca. 2500 v. Chr.) bis zur Römischen Kaiserzeit (3. Jh. n. Chr.) ab, wobei manche Epochen nur durch sehr magere oder gar keine Schriftquellen vertreten waren. Möller hatte geplant, diese Listen als Vorarbeiten für eingehende Untersuchungen zur Buchschrift, später auch zur Geschäftsschrift, zu verwenden und stetig zu erweitern. Aufgrund seines

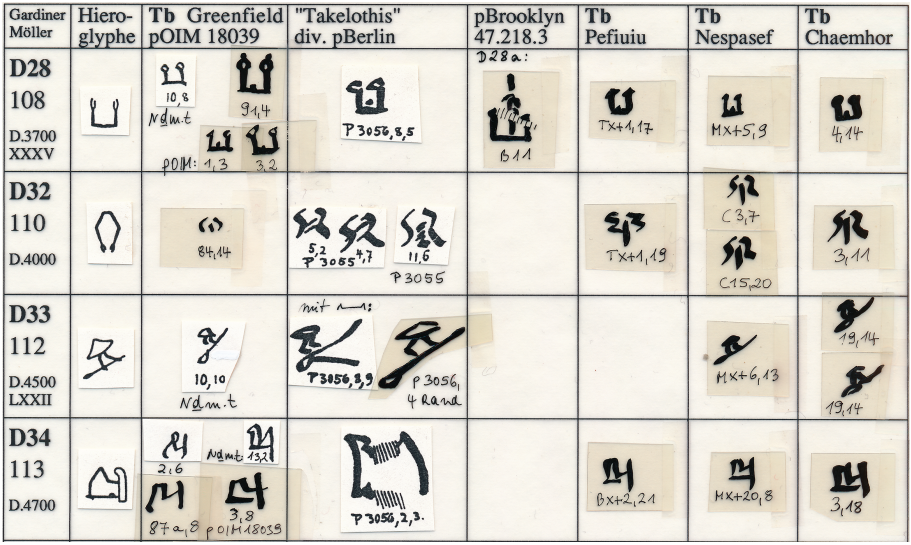


Abbildung 3: Ausschnitt aus einer traditionell erstellten paläographischen Liste, Arbeitsmaterial für die spätere Publikation: Verhoeven 2001, 118.

frühen Todes wurde dieses Unternehmen jedoch nicht vollendet und über 70 Jahre lang kam es kaum zu wesentlichen Fortschritten auf diesem Gebiet. Die jüngeren Teilpaläographien (Goedicke 1988; Verhoeven 2001; Allen 2002; Lenzo 2011, um nur einige zu nennen), halten sich bis heute an das Prinzip von Möller, allerdings in der Anordnung der *Sign-list* von Gardiner (Abb. 3) und mit diversen Erweiterungen und Kommentaren, im besten Fall unter Angabe der Strichfolge (Abb. 4).

Paläographische Listen, denen möglichst gut datierte Quellen zugrunde liegen, ermöglichen den Zeichenvergleich mit weiteren und zunächst undatierten Quellen, wodurch eine zeitliche Einordnung, wenn nicht sogar Zuschreibung an einen bestimmten Schreiber oder eine regionale Herkunft nahegelegt werden können. Die Kapazität einer gedruckten Liste ist selbstverständlich begrenzt und viele Paläographien enthalten nur ausgewählte Zeichen aus dem ca. 500 bis 600 Zeichen umfassenden Inventar des Hieratischen. In der Forschung existieren inzwischen zahlreiche kleinere oder größere Sammlungen von Zeichenbeispielen, manchmal nur eines einzigen Manuskripts, die idealerweise zusammengeführt werden sollten. Ein dynamisches Archiv ist daher wünschenswert und würde es ermöglichen, Recherche und Analyse bedeutend auszuweiten und neue Visualisierungsarten zur Verfügung zu stellen.

Vor über 40 Jahren forderte Georges Posener bereits einen *nouveau Möller* (Posener 1973) und formulierte die Aufgaben einer zukünftigen Erforschung des Hieratischen.

Individual Signs

A1 	I (2), (vo. 1) II (1), (29), (33), (2) III (8), (vo. 3), (3), (4) IV (3), (4), (2), (vo. 2) V (29), (25), (46) — columnar text only VII (15) See also Ligatures (N35).
A1* 	I (8) — as det. with B1* II (vo. 6) — as det. with B1* and in the account; inserted as 1s suffix in 29 V (13), (15), (30), (45) — in accounts VI (3), (7), (15) VII (7) P (9), (6), (19) Fr. (B); (C)
A2 	I (vo. 7); (vo. 17, in restricted space) II (5) III (1, 75%) — calligraphic, for A17 IV (4) V (26)
A15 	P (1)
A15* 	VII (2)

Abbildung 4: Ausschnitt aus einer digital erstellten Paläographie, Allen 2002, 193

Ein zentraler Punkt sollte dabei sein, dass das gesamte Formenspektrum eines Manuskripts aufgenommen werden sollte, und zwar unter Angabe der Häufigkeit der Zeichenformen, die vom Einfachen zum Komplexen angeordnet sein sollten. Eine Umsetzung gelang bislang jedoch nur in geringem Ausmaß bei vereinzelten Texteditionen.

Gegen Ende des 20. Jahrhunderts gab es die ersten Ansätze für eine digitale Aufbereitung und Analyse von hieratischen Handschriften bzw. Paläographien (Gosline 1999; dazu kritisch und mit eigenen grundsätzlichen Überlegungen: Van den Berg und Donker van Heel 2000). In vorliegender Publikationsreihe berichtete Stephen Quirke (2010) über das Projekt *The Lahun Papyri*, in dem eine computerunterstützte Paläographie für die zahlreichen Papyri aus der ägyptischen Siedlung Lahun zur Anwendung kommen sollte.

Bei allen technischen Möglichkeiten ist festzuhalten, dass auch eine digitale Paläographie sich an den Forderungen Poseners orientieren sollte, um sowohl den etablierten als auch den neuen Methoden und Forschungsfragen an die kursiven Handschriften gerecht werden zu können. Die aus den digitalen Methoden resultierenden Perspektiven, aber auch die hermeneutischen Herausforderungen, wurden 2011 auf der ersten Tagung *Ägyptologische „Binsen"-Weisheiten* präsentiert (Gülden 2016; Verhoeven 2015a, 51-54). Seit 2015 arbeitet nun das langfristig angelegte Mainzer Akademievorhaben *Altägyptische Kursivschriften (AKU)* an einer digitalen Paläographie in Form eines relationalen Datenbanksystems, um eine Grundlage für die systematische Analyse des Hieratischen und der Kursivhieroglyphen unter verschiedensten Gesichtspunkten liefern zu können. Das interdisziplinäre Projekt soll es ermöglichen, neue Tendenzen in den digitalen Geisteswissenschaften aufzugreifen und gegebenenfalls mitzubestimmen. Unabdingbar für ein solches Projekt sind der internationale Austausch von spezialisierten Ägyptologen (Verhoeven 2015a; 2015b) sowie die globale Zusammenarbeit bei der Erstellung von Editionen handschriftlicher Manuskripte, der Extraktion der Zeichenrepertoires sowie der erforderlichen digitalen Daten.

2 Metadaten und Datenmodell

Ein Metadatenmodell für eine quantitative Auswertung sollte flexibel sein und sowohl einfache als auch komplexe Beschreibungen aufnehmen können. Raum- und Zeitdaten sollten genauso abgebildet werden können wie Objektmetadaten, bibliographische Metadaten, Daten zur Beschreibung von Phänomenen der Schrift, des Schreibens und des Beschriftungsvorgangs. Die im Projekt konzipierte paläographische Datenbank kann für die Erfassung der Metadaten zu Textträgern auf eine Datenbankstruktur des Projektes *Trismegistos* zurückgreifen. Das von *Trismegistos* bereitgestellte Online-Portal liefert umfangreiche Metadaten papyrologischer und

epigraphischer Ressourcen aus der Zeit von 800 v. Chr. bis 400 n. Chr. (Gülden 2008). Neben Angaben zu Sprache, Textkategorie und Schrift sind darin auch die Datierung einer Textniederschrift,[2] der Herkunftsort des Textträgers,[3] die aufbewahrende Sammlung, im Text erwähnte Personen und deren Funktionen sowie bibliographische Hinweise enthalten. Informationen über den Schriftträger und das verwendete Schreibwerkzeug sind in *Trismegistos* ebenfalls, wenngleich knapp, aufgenommen worden. Da jedoch die individuelle Ausprägung der Schrift immer auch durch die Beschaffenheit des Textträgers beeinflusst wird, werden im AKU-Projekt zusätzlich wichtige Eckdaten zum jeweiligen Objekt erfasst. Dazu zählen neben Angaben zu Gattung, Material, Herkunft, Standort und Datierung auch Maßangaben, Angaben zu Erhaltungszustand, Wiederverwendung und zu äußeren Merkmalen (Werklayout). Außerdem werden Informationen zur Beschaffenheit der Oberfläche, zu Abfolge und Positionierung der Beschriftung sowie zur Drehung des Objekts während der Beschriftung aufgenommen. Im Zentrum stehen die Metadaten zu den Eigenschaften der Einzelzeichen und Zeichengruppen, von denen manche in enger Verbindung mit ihrem Trägerobjekt stehen, z. B. Schreibrichtung und -verlauf, Schreibwerkzeug oder Notationsart, also die Angabe, ob die Beschriftung mit Tusche oder in Ritzung vorgenommen wurde. Schließlich gehören Angaben über die Position auf dem Schriftträger[4] in diese Kategorie. Andere Metadaten beziehen sich auf die Schriftökonomie, z. B. Strichanzahl, Zeichenreduzierung[5], Zeichenverknüpfung (Ligatur, Teilligatur) oder auf die Gestalt(-ung) der Zeichen[6].

Eine erste Zusammenstellung der Einheiten für ein Metadatenmodell gliedert sich in zwei Hauptzweige. Der eine Zweig wird alle erwähnten Metadatenkategorien zum Textzeugen enthalten, die von *Trismegistos* übernommen worden sind. Der andere Zweig besteht aus aufeinander aufbauenden Abschnitten, die paläographisch relevante Einheiten wiedergeben. Diese Einheiten sind gruppiert von der größten paläographischen Einheit – dem beschrifteten Objekt[7] – über Untereinheiten – Text(-block)/Kolumne, Zeile, Zeichengruppe – hin zur kleinsten Texteinheit – dem Einzelzeichen. Der XML-Quellcode, der den Vektorgraphiken hinterlegt ist, enthält unter anderem bereits präzise Angaben zur Zeichengröße (Attribute *width* und *height* in px), die für eine Auswertung brauchbar sind. Die Umrisse der Schriftzeichen bzw. – falls erkennbar – der einzelnen Strichfolgen werden zudem im Pfaddatenelement erfasst. Die Metadaten zu den Texten und Textzeugen sollen unmittelbar mit den Metadaten zu den einzelnen Schriftzeichen verknüpft werden können. Auch sollte es

---

2 Die Urheber der Datierung eines Textzeugen, d. h. einer vorliegenden Handschrift, werden benannt und ggf. divergierende Vorschläge verzeichnet.

3 Die Herkunft von Text und Textträger muss nicht dieselbe sein, z. B. wenn derselbe Text auf unterschiedlichen Schriftzeugen und ggf. aus verschiedenen Orten und Zeiten überliefert ist.

4 Bei Papyrus insbesondere recto/verso; ansonsten vor allem Kolumne, Zeile etc.

5 Vollform [regulär, Variante] oder Kurzform [regulär, Variante].

6 Tuschefarbe, Strichfolge, Morphologie, Höhe und Breite, Grad der Zerstörung etc.

7 Einer Art Erweiterung der Einheit *Textträger*, die paläographisch relevante Daten aufnehmen wird.

möglich sein, Verbindungen zwischen Vollformen und Abkürzungen, Regelformen und Varianten oder Einzelzeichen und Zeichengruppen zu ziehen. Die Daten sollen über ein Metadatenschema für den regelmäßigen Export in TEI-konformes XML überführt werden. Alle Informationen zu den Schriftzeichen werden im `<teiHeader/>` aufgeführt. Der Aufbau des Headers orientiert sich an den vier Elementen des *Thot Data Model Object*, *Document*, *Witness* und *Text* (Polis und Razanajao 2016, 26-7).

Der grobe Aufbau eines TEI-Dokuments kann folgendermaßen aussehen:

`<teiHeader>`

- Informationen zur paläographischen Entität (Einzelzeichen: Hieratogramm, Kursivhieroglyphe; Zeichengruppe [Kombination und Ligatur])
- Informationen zum beschrifteten materiellen Artefakt, z. B. Papyrusfragment (»Object«)
- Informationen zum Textträger oder idealisiertem Schreibraum (»Document«)
- Informationen zum Textzeugen, z. B. Papyrus als Beleg für Lehre des Amenemhet (»Witness«)
- Informationen zum Text, z. B. Lehre des Amenemhet als Werk eines Autors (»Text«)

`<facsimile>`

- Links und Informationen zu den Bildern der Textzeugen (intern oder extern)
- Links (und Informationen) zu den Bildern der Schriftzeichen

Die verschiedenen Kategorien von Forschungsdaten wie Orte,[8] Namen,[9] zeitliche Angaben sollen soweit als möglich durch Normdaten und mit Hilfe von externen Thesauri nach etablierten bibliothekarischen Standards erschlossen werden.

Die Präsentationsformen der Hieratogramme sind bei einer digitalen Paläographie variabel. Nebeneinander können verwendet werden:

a) Ausschnitte aus einem Scan oder digitalen Foto des originalen Schriftzeugen inklusive umgebender Zeichen und der Oberfläche des Beschriftungsmaterials,
b) Faksimiles, d. h. Umzeichnungen, in denen die Einzelstriche schwarz ausgefüllt und nicht separat gekennzeichnet sind – wie in traditionellen Paläographien – und
c) Faksimiles, d. h. Umzeichnungen, in denen die Einzelstriche eines Zeichens nur mit ihren Umrisslinien wiedergegeben werden, sodass die Strichfolge erkennbar ist.

---

[8] *Trismegistos* arbeitet mit Georeferenzierungen von *Pleiades* (darüber auch *Pelagios*) und *GeoNames*. Alle geographischen Daten werden über eine Schnittstelle zu Google Maps angezeigt. Geographische Koordinaten können durch Geotagging hinzugefügt werden.

[9] Derzeit gibt es in *Trismegistos* etwa 34.500 verschiedene Personennamen, die prosopographisch ausgewertet werden können.

3 Zeichenlisten, Zeichenkodierung und Zeichenbeschreibung

Wie zu Beginn dargestellt, orientieren sich heutige Paläographien zumeist an der Zeichenliste Gardiners (Gardiner [3]1973, 438–548). Allerdings finden nicht alle Hieratogramme eine hieroglyphische Entsprechung in seiner *Sign-list*, und in den zwischenzeitlich publizierten Paläographien wurden vielfach neue Nummerierungen vergeben. Eine digitale Paläographie bietet den Vorteil, dass alle Nummerierungssysteme nebeneinander erfasst und flexibel erweitert werden können sowie nach eigenen Anforderungen recherchierbar sind. Im Folgenden werden Methoden der Kodierung und Beschreibung altägyptischer Schriftzeichen im digitalen Medium skizzenhaft beleuchtet.

Die erwähnten Angaben aus der Datenbank können zukünftig mit weiteren Angaben zu den einzelnen Schriftzeichen angereichert werden. Bei der Erarbeitung eines Informationsmodells für die Überführung dieser Angaben in einen digitalen Code empfiehlt es sich, von einem mehrgliedrigen System auszugehen, in dem sich die Klassifizierung eines Zeichens widerspiegelt.

Stéphane Polis und Serge Rosmorduc haben eine dreistufige Systematik vorgeschlagen, die es erlaubt, ein Schriftzeichen innerhalb der Hierarchie verschiedener Ausprägungen zu verorten (Polis und Rosmorduc 2013, 64–5). An oberster Stelle steht die ideale abstrakte Einheit des Schriftsystems (*graphème*), welche bestimmte minimale funktionale Eigenschaften aufweist, in der Mitte der Hierarchie steht die Klasse (*classe*), welche alle graphischen Varianten umfasst, die bildhafte Modifikationen gegenüber dem Graphem aufweisen, jedoch dieselben Funktionen besitzen. Auf der untersten Ebene steht die Form (*forme*), die den geringsten Abstraktionsgrad besitzt. Hier sind all jene graphischen Darstellungen zu finden, die Ausprägungen unterschiedlicher handschriftlicher Formvarianten innerhalb einer Klasse sind.[10] Im Projekt *Altägyptische Kursivschriften* wird eine Kodierung entwickelt, die sich an den hieratischen Zeichen mit ihren reichen Zeichenformen orientiert. Dafür wurden drei Elemente der Kodierung definiert (vgl. Abb. 5):[11]

1. Hauptzeichen (rot)
2. Formklassen (grün)
3. Verwendung in einer Zeichengruppe oder Ligatur (blau)

---

[10] Ein ähnliches System wurde von Meeks (2013) vorgeschlagen. Siehe zu einem Kommentar und Vergleich Meeks 2015. In diesem Rahmen ergibt sich leider nicht die Möglichkeit, beide Vorschläge im Detail zu besprechen.

[11] Die drei Stufen werden in einer Publikation (Van der Moezel, in Vorbereitung für Hieratic Studies Online) weiter erklärt werden. Das Hauptzeichen ist der ersten Stufe von Meeks (2013) vergleichbar (d. h. die Hauptzeichen haben keine feste, konkrete Repräsentation in der Schrift). Das System soll parallel zu anderen Nummerierungssystemen verwendet werden.

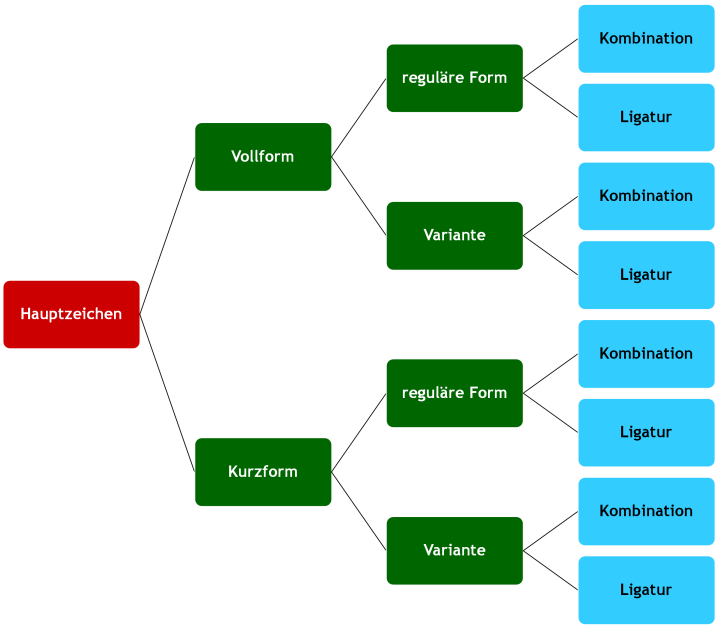


Abbildung 5: Vereinfachtes und vorläufiges Klassifikationsschema der geplanten AKU-Zeichenkodierung (vgl. Anm. 11)

Die oberste Bezugsgröße (rot) gibt die allgemeine Einordnung der Hauptzeichen des hieratischen Repertoires an. Das zweite Element hat zwei Bestandteile (grün) und gibt an, ob ein Zeichen in einer Vollform oder in einer abgekürzten Form, und darin wiederum in einer regulären Form oder in einer Variante ausgeführt ist. Die ersten beiden Elemente spiegeln also eine erste Formanalyse des Zeichens. Das dritte Element (blau) ordnet Zeichengruppen, die entweder aus solitären Zeichen bestehen oder in einer Ligatur miteinander verbunden sind, unter ihrem Hauptzeichen ein. Die Einordnung der Zeichengruppen richtet sich nach der Formklasse des Hauptzeichens innerhalb der Zeichengruppe. Das geschilderte Modell orientiert sich zwar grundsätzlich an den Hieroglyphen, jedoch erlaubt es aufgrund des Bezuges zwischen Hieroglyphe und hieratischem Schriftzeichen zumindest eine grobe formale Klassifikation von Hieratogrammen, die zudem in eine analysierbare Form, z. B. ein hierarchisch geschachteltes XML-Dokument überführt werden kann.

Peter Stokes war der Meinung, dass eine konventionelle hierarchische XML-Struktur für eine komplexe Schrift- und Buchstabenbeschreibung nicht ausreicht. Er schlug deshalb für mittelalterliche Handschriften ein alternatives Modell vor und unterschied zwischen den Kategorien *Schrift* (als imaginäres Konzept) und *Handschrift*,

also dem, was der Schreiber in physischer Form zu Papier bringt. Sein Modell bestand zu Beginn aus drei Entitäten. Die erste Kategorie, die einen Buchstaben als abstrakte Einheit definiert, umschreibt er mit dem Begriff *character*. Diese Einheit ist in etwa mit dem von Polis und Rosmorduc erwähnten *graphème* vergleichbar, steht aber letztlich zwischen *graphème* und *allograph*, der zweiten Kategorie bei Stokes (Stokes 2012).[12] Die Bezeichnung *allograph* umschreibt die besondere Art der Ausführung eines Schriftzeichens, referiert also auf seine Morphologie und kann in die Nähe der *classe* bzw. der *forme* bei Polis und Rosmorduc gestellt werden. Später fügte Stokes noch zwei weitere Entitäten hinzu: *idiograph* (konkrete Manifestation eines Allographen durch einen individuellen Schreiber) und *graph* (physische Instanz, konkrete Realisierung eines *character* durch die Hand eines Schreibers).[13] Die dritte Kategorie ist mit *component* betitelt. Diese bezeichnet die Grundbestandteile von Buchstaben, rekurriert also mehr auf die stilistische Ebene. Komponenten können ihrerseits wieder bestimmte Eigenschaften oder Ausprägungen (*features*) besitzen[14] (Stokes 2011, Part I; Stokes 2012). Die Kategorien seines Begriffsmodells verband er durch verschiedene Relationentypen. Alle Entitäten und ihre Relationen zueinander sind schließlich in ein erweitertes Klassendiagramm eingeflossen (Stokes 2011, Part IV; Stokes 2012, Abb. 2). Stokes' Modell ist imstande, nicht nur Buchstabenformen aufzunehmen, sondern ebenso spezielle Merkmale, die bei mehreren Buchstabenformen auftreten.

In der Paläographie-Datenbank des AKU-Projekts werden die Einzelzeichen und Zeichengruppen selbst umfassend beschrieben, und zwar vor allem in Bezug auf folgende Parameter: Materialität (i. e. Notationsart), Größe, Farbe, Anordnung auf dem Schriftträger, Schreibrichtung, Zeichenform, Strichfolge, Formklasse (Regelform oder Variante). Eine Beschreibung der formalen Eigenschaften kann sowohl für die Hauptklasse eines Hieratogramms als auch für seine Formklasse vorgenommen werden. An der Ausgestaltung der handgeschriebenen Zeichen lässt sich beispielsweise der Grad der Kursivität, ihre Formentwicklung von bildhaft zu abstrakt oder auch ihre Nähe zu den Hieroglyphen ablesen. Für einen paläographischen Vergleich sollten die Formvarianten durch Markup kenntlich gemacht werden. Im von der *EpiDoc Collaborative* zusammengestellten Subset der TEI-Guidelines sind einige Elemente für die Beschreibung der äußeren Gestalt von Buchstaben und Symbolen enthalten, allerdings sind diese bislang kaum für eine detaillierte Formbeschreibung von Schriftzeichen verwendbar.[15] Zudem orientieren sie sich ausschließlich an Alphabetschriften, weshalb sie für die altägyptischen Kursivschriften ungeeignet sind. Das Modul *Characters,*

---

[12] »Thus the grapheme <a> has (at least) two characters. 'capital' **A** and 'small' **a**. The second of these has many allographs, one of which is Insular a (…)«.

[13] »So allographs function at the level of script, and idiographs (as well as graphs) at the level of scribal hands«.

[14] »(…) thus a descender may be straight or curved, long or short, and so on.«

[15] Die Empfehlungen der *EpiDoc*-Gemeinschaft werden für die Transkription von Schriftdokumenten des Altertums eingesetzt und sind inzwischen von antiken Inschriften auf Papyri und Manuskripte erweitert worden.

Glyphs and Writing Modes (gaiji) in den TEI-Guidelines beinhaltet Elemente für die Beschreibung von Schriftzeichen, insbesondere in den Abschnitten *Markup Constructs for Representation of Characters and Glyphs und Annotating Characters*. Für die Beschreibung von Handschrift kommen beispielsweise die Elemente `<scriptDesc>` und `<handDesc>` im Modul *Manuscript Description* (`<msDesc>`) in Frage. Eine Beschreibung kann hier im Prosatext über `<p>` oder im Header über die Elemente `<scriptNote>` bzw. `<handNote>` mit Verweis auf eine `xml:id` und Attributen erfolgen, jedoch gibt es kein Vokabular, um bestimmte Zeichen und ihre Merkmale in einer formalisierten Art und Weise beschreiben zu können. Für die Auszeichnung der großen Varietät in der Morphologie nicht-alphabetischer handgeschriebener Zeichen erscheinen die Richtlinien also unzureichend. Wohl auch aus diesem Grund hat sich die Arbeitsgruppe *ENcoding COmplex Writing Systems (ENCOWS)* aus Vertretern unterschiedlicher Fachgebiete gebildet, welche sich mit Fragen rund um die Auszeichnung komplexer, nicht-alphabetischer Schriftsysteme auseinandersetzt.

Darüberhinaus bietet das *Unicode Consortium* Lösungsvorschläge an, die ganz allgemein auch für Kodierungsformen hieratischer Schriftzeichen dienlich sein könnten. Für Einzelzeichen liefert der *Unicodeblock CJK-Striche* ein Beispiel, indem er vereinheitlichte Strichtypen der *Character Description Language (CDL)* zusammenfasst. Hierbei handelt es sich um eine Zeichenbeschreibungssprache für asiatische Sprachen. Für Zeichengruppen kann die Beschreibung von bislang unkodierten Schriftzeichen, die sich aus zwei oder drei bereits kodierten Zeichen zusammensetzt, eine Vorlage sein. *Unicode* verwendet eine ideographische Beschreibungssequenz, die das Schriftzeichen als stilisiertes Bild begreift. Mithilfe des *Unicodeblocks Ideographische Beschreibungszeichen* wird angegeben, wie die Schriftzeichen kombiniert werden können. Die Aufteilung in einzelne Segmente visualisiert das Vorgehen bei der Beschreibung.

Im Falle der hieratischen Schrift wäre zunächst eine eigene Zeichenbeschreibung nach formalen Kriterien sinnvoll, die bislang ein Desiderat darstellt. Äußere Merkmale und distinktive Züge der Hieratogramme sollten vorzugsweise in normierter und standardisierter Form festgehalten werden. Diese Forderung ist nicht neu, denn Überlegungen zu einer eindeutigen und standardisierten Beschreibung von Handschrift existieren in der paläographischen Forschung bereits seit dem 18. Jahrhundert (Stokes 2011, Part I für Alphabetschrift). Léon Gilissen nennt sechs Kriterien für eine objektive Schriftbeschreibung (Bromm 1999, 27f.):

- Schriftwinkel (zwischen Schreibgerät und Zeile oder Grundstrich bzw. Haarstrich und Zeile)
- Modul (absolute oder relative Höhe und Breite des durchschnittlichen Schriftzeichens)
- Gewicht (berechnet aus Schriftwinkel, Breite des Schreibgeräts und Größenverhältnis der Buchstaben)

- Duktus (Strichreihenfolge und –richtung)
- Morphologie oder Grundgestalt (Aussehen und Form der Zeichen)
- Stil (nicht messbarer uniformer Charakter der Zeichen)

Peter Stokes unterscheidet zwei unterschiedliche Herangehensweisen an die Beschreibung von Schriftzeichen, den morphologischen Ansatz (*morphological*) und den stilbasierten Ansatz (*style-based*). Der morphologische Ansatz ist der Zeichenform als Ganzer gewidmet. Bei den verwendeten Begriffen kann eine Hierarchie von einer Grob- zu einer Feinbeschreibung gebildet werden. Der stilbasierte Ansatz bezieht sich auf den Gesamteindruck des Schriftbildes und auf bestimmte Komponenten der Schrift, die dafür kennzeichnend sind (Stokes 2011, Part I). Im Projekt *Stefan George Digital. Eine typographisch erschlossene Digitale Edition* wird eine mikrotypographische Modellierung erarbeitet, welche die Aspekte *Form* und *Stil* (ausgedrückt durch Einflüsse nicht-lateinischer Schriftarten wie dem Griechischen) sowie *semantische Funktion*, d. h. die Verwendung von Schriftarten in bestimmten semantischen Kontexten, berücksichtigt (Neuber 2016). Die beiden letzten Kategorien lassen sich von der Typographie auf die hieratische Handschrift allerdings wohl nicht ohne weiteres übertragen. Das Projekt *Altägyptische Kursivschriften* konzentriert sich daher zunächst auf die reine Formbeschreibung der kursiven Schriftzeichen.

Hilfreich wäre der Einsatz eines kontrollierten Vokabulars, welches bereits von verschiedener Seite gefordert worden ist (Gülden 2016; Polis und Rosmorduc 2013, 62). Zuvor festgelegte Begriffe für bestimmte Eigenschaften können in eine große Taxonomie einfließen, die eine hierarchische Auffächerung von der Grob- zur Detailbeschreibung erlaubt. Für eine Grobbeschreibung der Hieratogrammform können Begriffe Vorbild sein, wie sie Gardiner einst in seiner *Sign-list* für Hieroglyphenformen verwendet hat (z. B. *tall narrow signs* oder *low broad signs*). Eine Möglichkeit der paläographischen Schriftbeschreibung im Detail ist die Zerlegung der Zeichen in ihre Einzelelemente, die mit einem allgemeinen Terminus versehen und anschließend mit näherer Spezifikation weiter aufgefächert werden (beispielhaft Tabelle 1). Die einzelnen Elemente und ihre Spezifizierungen können für die Beschreibung anschließend zusammengesetzt werden. Außerdem können zusätzlich Angaben zum Verlauf dieser Elemente gemacht werden (*von links nach rechts*, *von oben nach unten* usw.) (Bromm 1999, 22; 23 Abb. 1; 24 Abb. 2 für die Beschreibung von Buchstaben).

Ein solches Beschreibungssystem ist vergleichbar mit der Systematik, welche im Projekt *Relationen im Raum (RIR)* für die Spezifizierung von Teilelementen bei jüdischen Grabsteinen erarbeitet wurde. Auch dort hat man komplexe Formen hierarchisch in Teilformen aufgegliedert und so weitere Hierarchieebenen geschaffen, die bei einer Datenbankabfrage schrittweise abgearbeitet werden können (Gietz et al. 2016, 12–15). Ob eine entsprechende Methode, die für Architekturbestandteile entwickelt wurde, auch für die Beschreibung kursiver Schriftzeichen in Frage kommt, muss erprobt

Einzelelement: Typ	Typ: Form	Form: Ausführung
Bogen	abfallend	-
	aufsteigend	konkav nach rechts
		konkav nach links
	horizontal	nach oben geöffnet
		nach unten geöffnet
Wellenlinie	geneigt	nach rechts
		nach links
	horizontal	-
	vertikal	-

Tabelle 1: Möglichkeiten der detaillierten Zeichenbeschreibung

werden. Das vorgegebene Vokabular könnte für eine Anreicherung des Quellcodes in Frage kommen, und zwar über ein Element `<object>` und Attributen, deren Werte den festgelegten Begrifflichkeiten folgen, etwa in der Form:

```
<object invnr="00014" category="Einzelzeichen" type="Hieratogramm">
  <object category="Einzelelement" type="Bogen" form="aufsteigend">
    <object category="Ausführung" form="konkav nach rechts"/>
  </object>
  <object category="Einzelelement" type="Wellenlinie" form="geneigt">
    <object category="Ausführung" form="nach rechts"/>
  </object>
</object>
```

Schließlich könnte auch die Anordnung einzelner Schriftzeichen, etwa in einer Zeichengruppe, mit Hilfe einer speziellen Kodierung umschrieben werden, die für Hieroglyphen erstmalig im sogenannten *Manuel de Codage (MdC)* (Buurman et al. 1988) definiert worden ist und ursprünglich für die korrekte Umsetzung der Hieroglyphen bei der Eingabe am Computer gedacht war (Rosmorduc 2015, 4). Um auszudrücken, dass zwei Zeichen mit Gardiner-Nummer sich in derselben Zeile befinden, wird ein Asterisk (*) als Trenner verwendet. Ein Zeilenbruch zur zweiten (unteren) Ebene der Zeichengruppe kann mit einem Doppelpunkt (:) angegeben werden, z. B. Q3*X1:N1. Dieses System ist vergleichbar mit dem Standard zur Wiedergabe der Lesefolge einzelner Maya-Hieroglyphen in einem Hieroglyphenblock, »nach dem (…) nebeneinander stehende Zeichen durch einen Punkt, übereinander stehende durch einen Doppelpunkt getrennt werden.«, z. B. T1:257.1:624:178[16] (Maier 2015, 17).

[16] Die Nummerierung der Maya-Hieroglyphen erfolgt nach der Zusammenstellung im Katalog von Eric Thompson aus dem Jahr 1962 (Thompson 1962).

4 Verfahren der Schriftanalyse

Quantifizierbare Merkmale von Schrift können eingesetzt werden, um verschiedene Schriftzeichen oder sogar ganze Handschriften miteinander zu vergleichen. Für eine quantitative Auswertung paläographischer Daten sind vor allem metrische Angaben interessant. In der Hieratistik gehören zu dieser Kategorie die Kolumnenzahl, Zeilenzahl, Strichzahl bzw. Zeichenzahl pro Zeile oder die Häufigkeit des Aufnehmens frischer Tusche (*dipping*; vgl. z. B. Allen 2002, 227–242; Verhoeven 2017, 64-66) sowie die Zeichenhäufigkeit im Text. Aus der Paläographie des Mittelalters sind weitere metrische Kategorien bekannt, die auch für hieratische Quellentexte eingesetzt werden können: Aus den Maßen zur Zeichengröße lassen sich Relationen berechnen, wie z. B. das Verhältnis von Höhe und Breite einzelner Zeichenformen (*Modulus*; Stokes 2009, 313). Für die Berechnung eines Durchschnittsbuchstabens werden möglichst viele Zeichen in Höhe und Breite vermessen, die Werte addiert und der jeweilige Betrag anschließend durch die Anzahl der Probanden geteilt. Die durchschnittliche Zeilenhöhe auf einer Seite kann durch die Höhe des Schriftspiegels in Millimeter, dividiert durch die Zeilenzahl, berechnet werden. Das Verhältnis zwischen der so berechneten Zeilenhöhe und der durchschnittlichen Zeichenhöhe gibt Aufschluss über die durchschnittliche Höhe einer Schrift; dividiert man die durchschnittliche Zeichenbreite durch die Zeilenhöhe erhält man ein Maß für die Schriftdichte (Bromm 1999, 27).[17] Auch andere Einheiten wie der Winkel der Strichdicke oder der Winkel der Strichneigung können für eine Schriftanalyse erhoben werden. Ähnlichkeiten und Unterschiede in Schriften können auf diese Weise gemessen und errechnet werden. Auch lässt sich zeigen, dass die Gestaltung von Schriftzeichen gewissen Gesetzmäßigkeiten unterliegt, z. B. kann überprüft werden, ob ihre Verteilung in Texten vom Komplexitätsgrad abhängig ist.

Für die Analyse paläographischer Daten können allgemein Verfahren aus der quantitativen Linguistik zur Anwendung kommen. In den letzten Jahren ist das Interesse an der Untersuchung von graphischen Symbolen und Schriften auf diesem Gebiet gewachsen und das nicht ohne Grund. Für das Begriffspaar *Type – Token* gibt es eine Entsprechung bei den hieratischen Schriftzeichen. Der Menge der Wortformen, die ein und dieselbe sprachliche Einheit (Wort) in einem Text darstellen (*Type*), entspricht demnach die Menge der vorkommenden Zeichenformen in einem Papyrus. Die kleinste sprachliche Einheit (*Token*) findet ihre Entsprechung in jeder einzelnen einmalig

[17] Eine andere Methode zur Bestimmung der Schriftdichte kann durch die Division der beschriebenen Fläche durch die Anzahl der Zeichen innerhalb der beschriebenen Fläche (Zeilenzahl multipliziert mit der Zahl der Zeichen pro Zeile) erreicht werden. Es gilt dabei: Je kleiner der Wert desto größer die Schriftdichte. Einfacher lässt sich die Schriftdichte bestimmen, indem anhand der Größe der Zeichen der eingenommene Raum im Verhältnis zum freien Raum auf dem Schriftträger berechnet wird, siehe dazu den Beitrag von Matthew Driscoll in diesem Band. Driscoll bestimmt aber nicht die Schriftdichte, sondern die Textdichte, indem er an die Stelle der Zeichen Wörter setzt.

Paläographie	Linguistik
Menge der Zeichenformen	Type
einzelnes vorkommendes Zeichen	Token
Grundzeichen	unflektierte Wortform
Variante, Allograph	flektierte Wortform
Einzelzeichen	Einzelwort
Zeichengruppe	Wortgruppe
Ligatur	Wortverbindung

Tabelle 2: Entsprechungen zwischen paläographischen Bezeichnungen und linguistischer Sprachterminologie

vorkommenden Zeichenform. Der erwähnten Grundform eines Zeichens (*character*) kann ohne Probleme die Grundform eines Wortes gegenübergestellt werden; die Varianten oder Allographen (*classe*, *allograph*) entsprechen dann den flektierten Formen eines Wortes. Einzelzeichen können Einzelwörter entgegengehalten werden, wohingegen Zeichengruppen bzw. Ligaturen mit den Wortgruppen bzw. Wortverbindungen korrespondieren (Tabelle 2).

Wenn derartige Verfahren aber auf unterschiedliche alphabetische wie nichtalphabetische Schriftsysteme übertragen werden sollen, dann müssen sie möglichst applikabel sein, um einen generischen Status zu erhalten. Vielversprechend für paläographische Fragestellungen scheint die Anwendung von Frequenzanalysen zu sein. Hier bringt man die Schriftzeichen gemäß der Frequenz ihres Auftretens in eine Rangordnung und kann somit etwa auch ihre Vermehrung im Laufe der Zeit bestimmen. Eine Methode aus der Lexikostatistik, das sogenannte *Vector Space Modeling*, hat Simon Schweitzer vorgestellt, bei der es um die Ermittlung der Lebensspanne von Wörtern in altägyptischen literarischen Texten geht (Schweitzer 2013). Er geht davon aus, dass ein bestimmtes Vokabular innerhalb einer bestimmten Zeitspanne am häufigsten vorkommt und die zeitliche Einordnung von Texten spiegelt. Alle Wortformen der untersuchten Dokumente werden als Tokens in einem allgemeinen Index gesammelt. Anschließend wird für jedes Dokument ein numerischer Vektor gebildet, dessen Anzahl der Stellen der Anzahl aller Tokens im allgemeinen Index entspricht. Das Vorkommen jeder Wortform wird als Zahl an der entsprechenden Stelle im Vektor festgehalten, z. B. (1 | 2 | 1 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0). Alle so gebildeten Vektoren können dann in eine Tabelle übertragen werden, wobei die Spalten den einzelnen Textzeugen und die Zeilen den jeweiligen Stellen im Vektor entsprechen. In den Feldern werden die Häufigkeitsangaben vermerkt. Anschließend wird die Matrix normalisiert, so dass die Länge der Vektoren dem Wert 1 entspricht. Zum Schluss werden die Distanzen zwischen den Vektoren berechnet. Geringe Distanzwerte zwischen

Vektoren deuten auf eine hohe Ähnlichkeit der Dokumente hin. Aus dem Vergleich kann abgeleitet werden, dass einige Texte sich in ihrer Wortwahl ähnlicher sind als andere und daher in eine ähnliche Zeitstufe eingeordnet werden können. Ob das von Schweitzer beschriebene Verfahren von Wortformen auf hieratische Schriftzeichen übertragen werden kann, wird zu prüfen sein. Ein Problem ist jedenfalls, dass Texte möglicherweise zeitgleich eingeordnet werden können, was aber nicht zwangsläufig bedeutet, dass auch die Abschriften zeitgleich sind. Von literarischen oder religiösen Texten wurden in unterschiedlichen Abständen, die mehrere Jahrhunderte umfassen können, immer wieder Kopien angefertigt, z. B. gibt es Abschriften der Lehre des Königs Amenemhet etwa um 1500, 1200 und 600 v. Chr. (Verhoeven 1999).

Ein weiterer Ansatz möchte die Komplexität von Zeichen in Texten ermitteln und ihre Verteilung untersuchen, um beispielsweise den Grad der Simplifizierung von Hieratogrammen und Hieroglyphen zu errechnen. Eine Möglichkeit besteht darin, die Anzahl der Striche, aus denen sich ein Zeichen zusammensetzt, zu zählen (so wie dies beispielsweise für chinesische oder akkadische Schriftzeichen gemacht wurde). Eine weitere Methode stammt wiederum aus der quantitativen Linguistik: Man zerlegt eine graphische Einheit in kleinere Bestandteile. Den Schriftelementen (Punkte, gerade Linien und Bögen) und den Arten ihrer Verbindungen (z. B. *fließend*, *scharf* oder *kreuzend*) wird jeweils ein numerischer Wert zugewiesen. Alle Punktzahlen werden anschließend addiert. Für jedes Graphem kommt so ein bestimmter Wert heraus. Die Annahme: Je höher dieser Wert ist, desto höher ist auch die graphische Komplexität eines Zeichens (Meletis 2015, 76–79). Ina Hegenbarth-Reichardt und Gabriel Altmann stellen die Komplexitätswerte von 20 Hieroglyphen und den zugehörigen Hieratogrammen einander gegenüber. Aus den gewonnenen Werten errechnen sie jeweils den Durchschnittswert und die Stichprobenvarianz, um einen Indikator für die Streuung zu erhalten. Im Vergleich zwischen Hieroglyphen, Hieratogrammen und der Schriftart Courier wurde mathematisch nicht nur nachgewiesen, dass die Komplexität der Hieroglyphen diejenige von Hieratogrammen stark übertrifft, sondern auch, dass die hieratische Schrift und die Schriftart Courier nahezu identische Komplexitätswerte besitzen. Die Verallgemeinerung, dass sich der Komplexitätsgrad eines Hieratogramms automatisch nach dem Komplexitätsgrad der Hieroglyphe richtet[18], konnte hingegen nicht bewiesen werden (Hegenbarth-Reichardt und Altmann 2008, 108–112).

5 Schluss

In einer digitalen Paläographie kommen computerbasierte Methoden zum Einsatz, die den geübten Blick des Paläographen unterstützen und das Studium der materiellen Textträger sinnvoll ergänzen können. Bestimmte Fragen hinsichtlich der Materialität

[18] Die Annahme: je komplexer eine Hieroglyphe, desto komplexer auch ihre hieratische Simplifikation.

historischer Objekte lassen sich nach wie vor nur am dreidimensionalen Original überprüfen. Dies gilt insbesondere für Studien an Einzelobjekten. In diesem Punkt stößt digitales Forschen an Schriftdokumenten bisweilen an Grenzen. Der entscheidende Vorteil eines quantitativen Zugangs liegt in einer beschleunigten Erfassung der Schriftzeichen und in einer Be- bzw. Auswertung großer Materialmengen. Unser Beitrag ist der Frage nachgegangen, wie paläographische Informationen zu altägyptischen Kursivschriften mit digitalen Methoden bearbeitet werden können, die dem Paläographen neue Wege aufzeigen, ihn aber gleichzeitig dazu zwingen, sein Wissen strukturiert aufzubereiten, um es anschließend auf einer breiteren Datengrundlage überprüfen zu können. Für eine möglichst präzise und detailgenaue Beschreibung von hieratischen Schriftzeichen im digitalen Medium ist es wichtig und notwendig, das eigene Verständnis der einzelnen Schriftzeichen zu überdenken und eine Ergründung ihres ureigenen Wesens vorzunehmen. Für das Metadatenmanagement oder für bestimmte Verfahren der Auswertung kann auf Konzepte und Informationsmodelle aus anderen Disziplinen zurückgegriffen werden. Es lohnt sich im Umgang mit dem Digitalen einen Blick über den Tellerrand der eigenen Disziplin zu werfen. Im Idealfall stellen sich dabei Synergieeffekte ein, die den Horizont für die eigenen Fragestellungen erweitern können.

Bibliographie

Allen, James P. *The Heqanakht papyri*. New York: Metropolitan Museum of Art, 2002.

AKU: *Altägyptische Kursivschriften. Digitale Paläographie und systematische Analyse des Hieratischen und der Kursivhieroglyphen.* <http://aku.uni-mainz.de>.

Assmann, Jan. »Die ägyptische Schriftkultur.« In Günther, Hartmut und Otto Ludwig (Hrsg.). *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung* I. Berlin, New York (NY): de Gruyter, 1994. 472–492.

Bromm, Gudrun. »Neue Vorschläge zur paläographischen Schriftbeschreibung.« In Rück, Peter (Hrsg.). *Methoden der Schriftbeschreibung*. Sigmaringen: Thorbecke, 1999. 21–42.

Buurman, Jan et al. *Inventaire des signes hiéroglyphiques en vue de leur saisie informatique*. Paris: De Boccard, 1988.

ENCOWS: *ENcoding COmplex Writing Systems*. <http://ancientworldonline.blogspot.de/2016/06/encoding-complex-writing-systems-encows.html>.

EpiDoc: Epigraphic Documents in TEI XML. XML text markup for ancient documents, (Hrsg. Gabriel Bodard et al.). <http://epidoc.sourceforge.net>.

EpiDoc Guidelines: Ancient documents in TEI XML (Version 8). 2007–2016. <http://www.stoa.org/epidoc/gl/latest>.

Gardiner, Alan H. *Egyptian Grammar*. Oxford: Oxford University Press, 1927, ^{3}1973.

Gietz, Peter et al. *Relationen im Raum Visualisierung topographischer Klein(st)strukturen [RiR]. Schlussbericht.* 31. Januar 2016. <https://wiki.de.dariah.eu/display/RIRPUB/Poster+und+Publikationen#PosterundPublikationen-Abschlussbericht>.

Goedicke, Hans. *Old Hieratic Paleography*. Baltimore (MD): Halgo, 1988.

Gosline, Sheldon L. *Hieratic Paleography 1: Introductory Late Egyptian.* Warren Center, Pa.: Shangri-La Publications, 1999.

Gülden, Svenja A. »*Trismegistos*: An interdisciplinary portal of papyrological and epigraphical resources.« In Strudwick, Nigel (Hrsg.). *Information Technology and Egyptology in 2008. Proceedings of the meeting of the Computer Working Group of the International Association of Egyptologists (Informatique et Égyptologie), Vienna, 8–11 July 2008.* Piscataway (NJ): Gorgias Press, 2008. 17–28.

Gülden, Svenja A. »Ein ›*nouveau Möller*‹? Grenzen und Möglichkeiten. Ein Working Paper zum gleichnamigen Vortrag.« *Hieratic Studies Online* 1 (2016). <https://publications.ub.uni-mainz.de/opus/frontdoor.php?source_opus=55758>.

Gülden, Svenja A., Celia Krause und Ursula Verhoeven. »Digital Palaeography of Hieratic.« In Davies, Vanessa und Dimitri Laboury (Hrsg.). *Oxford Handbook of Epigraphy and Palaeography*. Oxford: Oxford University Press, (im Druck).

Hegenbarth-Reichardt, Ina und Gabriel Altmann. »On the decrease of complexity from hieroglyphs to hieratic symbols.« In Altmann, Gabriel, and Fan Fengxiang (Hrsg.). *Analyses of Script: Properties of Characters and Writing Systems.* Berlin, New York (NY): de Gruyter, 2008. 105–114.

Leach, Bridget und John Tait. »Papyrus.« In Nicholson, Paul T. und Ian Shaw (Hrsg.). *Ancient Egyptian Materials and Technology.* Cambridge: Cambridge University Press, 2000. 227–253.

Lenzo, Giuseppina. »Paleografia.« In Roccati, Alessandro (Hrsg.). *Magica Taurinensia. Il grande papiro magico di Torino e i suoi duplicati.* Rom: Gregorian & Biblical Press, 2011. 193–255.

Maier, Petra. *Die Erstellung eines TEI-Metadatenschemas für die Auszeichnung von Texten des Klassischen Maya. DARIAH-DE Working Papers 8.* Göttingen: DARIAH-DE, 2015. URN: urn:nbn:de:gbv:7-dariah-2015-1-6.

Meeks, Dimitri. »Dictionnaire hiéroglyphique, inventaire des hiéroglyphes et Unicode.« *Document numérique* 16.3 (2013). 31–44.

Meeks, Dimitri. »Linguistique et égyptologie: entre théorisation à priori et contribution à l'étude de la culture égyptienne.« *Chronique d'Égypte* 90 (2015). 40–67.

Meletis, Dimitrios. *Graphetik. Form und Materialität von Schrift.* Glückstadt: Verlag Werner Hülsbusch, 2015.

Möller, Georg. *Hieratische Paläographie. Die Aegyptische Buchschrift in ihrer Entwicklung von der fünften Dynastie bis zur Römischen Kaiserzeit I–III.* Leipzig: J. C. Hinrichs, 1909–1912. I–IV: Leipzig: J.C. Hinrichs, ^{2}1927–1936. Neudruck Osnabrück: Otto Zeller, 1965.

Neuber, Frederike. »Stefan George Digital: Exploring Typography In A Digital Scholarly Edition.« *Digital Humanities 2016: Conference Abstracts.* Kraków: Jagiellonian University und Pedagogical University. 637–639.

Parkinson, Richard B. und Stephen Quirke. *Papyrus.* London: British Museum Press, 1995.

Polis, Stéphane und Vincent Razanajao. »Ancient Egyptian texts in context. Towards a conceptual data model (The Thot Data Model — TDM).« *BICS* 59.2 (2016). 24-41.

Polis, Stéphane und Serge Rosmorduc. »Réviser le codage de l'égyptien ancien: Vers un répertoire partagé des signes hiéroglyphiques.« *Document Numérique* 16.3 (2013). 45–67.

Posener, Georges. »L'écriture hiératique.« In *Textes et langages de l'Égypte pharaonique, cent*

cinquante années de recherches I, 1822–1972. Kairo: Institut français d'archéologie orientale, 1973. 25–30.

Quirke, Stephen. »Agendas for Digital Palaeography in an Archaeological Context: Egypt 1800 BC.« In Fischer, Franz, Christiane Fritze und Georg Vogeler (Hrsg.). *Kodikologie und Paläographie im digitalen Zeitalter* 2. Norderstedt: Books on Demand, 2010. 279–294.

RIR: *Relationen im Raum.Visualisierung topographischer Klein(st)strukturen.* <https://wiki.de.dariah.eu/display/RIRPUB/RiR>.

Rosmorduc, Serge. »Computational Linguistics in Egyptology.« In Stauder-Porchet, Julie, Andreas Stauder und Willeke Wendrich (Hrsg.). *UCLA Encyclopedia of Egyptology*. Los Angeles (CA), 2015. <http://digital2.library.ucla.edu/viewItem.do?ark=21198/zz002jh4wt>.

Schweitzer, Simon D. »Dating Egyptian Literary Texts: Lexical Approaches.« In Moers, Gerald et al. (Hrsg.). *Dating Egyptian Literary Texts*. Hamburg: Widmaier Verlag, 2013. 177–190.

Stokes, Peter. »Computer-Aided Palaeography, Present and Future.« In Rehbein, Malte et al. (Hrsg.). *Kodikologie und Paläographie im digitalen Zeitalter*. Norderstedt: Books on Demand, 2009. 310–338.

Stokes, Peter. *Describing Handwriting Part I*. Working paper. London: King's College London, 2011. <http://www.digipal.eu/blog/describing-handwriting-part-i>.

Stokes, Peter. »Modeling Medieval Handwriting: A New Approach to Digital Palaeography.« In Meister, Jan Christof (Hrsg.). *DH2012: Book of Abstracts*. Hamburg, 2012. 382–385.

TEI-Guidelines: *Text Encoding Initiative, P5 Guidelines.* <http://www.tei-c.org/Guidelines/P5/>.

The Lahun Papyri. <http://www.ucl.ac.uk/museums-static/digitalegypt/lahun/papyri.html>.

Thompson, Eric. *A Catalog of Maya Hieroglyphs*. Oklahoma (OK): University of Oklahoma Press, 1962.

Trismegistos. An interdisciplinary Platform for Ancient World Texts and Related Information. <http://www.trismegistos.org>.

Van den Berg, Hans und Koenraad Donker van Heel. »A Scribe's Cache from the Valley of the Queens? The Palaeography of Documents from Deir el-Medina: Some Remarks.« In Demareé, Rob J. und Arno Egberts (Hrsg.). *Deir el-Medina in the Third Millenium A. D.*. Leiden: Nederlands Instituut voor het Nabije Oosten, 2000. 9–49.

Van der Moezel, Kyra. »On Signs, Lists and Standardization.« In Gülden, Svenja A., Kyra van der Moezel und Ursula Verhoeven (Hrsg.). *Ägyptologische „Binsen"-Weisheiten* III. Stuttgart: Franz Steiner Verlag, (in Vorbereitung).

Verhoeven, Ursula. »Von hieratischen Literaturwerken in der Spätzeit.« In Assmann, Jan und Elke Blumenthal (Hrsg.). *Literatur und Politik im pharaonischen und ptolemäischen Ägypten, Vorträge der Tagung zum Gedenken an Georges Posener 5.–10. September 1996 in Leipzig*. Kairo: Institut français d'archéologie orientale, 1999. 255–265.

Verhoeven, Ursula. *Untersuchungen zur späthieratischen Buchschrift*. Leuven: Peeters, 2001.

Verhoeven, Ursula (Hrsg.) [2015a]. *Ägyptologische „Binsen"-Weisheiten* I–II. *Neue Forschungen und Methoden der Hieratistik. Akten zweier Tagungen in Mainz im April 2011 und März 2013*. Stuttgart: Franz Steiner Verlag, 2015. <https://publications.ub.uni-mainz.de/opus/frontdoor.php?source_opus=54754>.

Verhoeven, Ursula [2015b]. »Stand und Aufgaben der Erforschung des Hieratischen und der Kursivhieroglyphen.« In Verhoeven, Ursula (Hrsg.).*Ägyptologische „Binsen"-Weisheiten*

I–II. *Neue Forschungen und Methoden der Hieratistik, Akten zweier Tagungen in Mainz im April 2011 und März 2013*. Stuttgart: Franz Steiner Verlag, 2015. 23–63. <https://publications.ub.uni-mainz.de/opus/frontdoor.php?source_opus=54754>.

Verhoeven, Ursula. *Das frühsaitische Totenbuch des Monthpriesters Chamhor C*. Unter Mitarbeit von Sandra Sandri. Basel: Orientverlag, 2017.

MEI Kodierung der frühesten Notation in linienlosen Neumen

Inga Behrendt, Jennifer Bain, Kate Helsen

Zusammenfassung

Das *Optical Neume Recognition Project* (ONRP) hat die digitale Kodierung von musikalischen Notationszeichen aus dem Jahr um 1000 zum Ziel – ein ambitioniertes Vorhaben, das die Projektmitglieder veranlasste, verschiedenste methodische Ansätze zu evaluieren. Die *Optical Music Recognition-Software* soll eine linienlose Notation aus einem der ältesten erhaltenen Quellen mit Notationszeichen, dem *Antiphonar Hartker* aus der Benediktinerabtei St. Gallen (Schweiz), welches heute in zwei Bänden in der Stiftsbibliothek in St. Gallen aufbewahrt wird, erfassen. Aufgrund der handgeschriebenen, linienlosen Notation stellt dieser Gregorianische Gesang den Forscher vor viele Herausforderungen. Das Werk umfasst über 300 verschiedene Neumenzeichen und ihre Notation, die mit Hilfe der *Music Encoding Initiative* (MEI) erfasst und beschrieben werden sollen. Der folgende Artikel beschreibt den Prozess der Adaptierung, um die MEI auf die Notation von Neumen ohne Notenlinien anzuwenden. Beschrieben werden Eigenschaften der Neumennotation, um zu verdeutlichen, wo die Herausforderungen dieser Arbeit liegen sowie die Funktionsweise des *Classifiers*, einer Art digitalen Neumenwörterbuchs.

Abstract

The *Optical Neume Recognition Project* (ONRP) is one branch of *Cantus Ultimus*, a research team overseen by Ichiro Fujinaga at McGill University in Montreal, Canada in association with the SIMSSA project (Single Interface for Score Searching and Analysis), also under Fujinaga's direction at McGill. *Cantus Ultimus* aims to use *optical music recognition* (OMR) technology to develop tools for searching plainchant manuscripts for musical information in much the same way that *optical character recognition* (OCR) is used to search text in digital environments. With thousands of digital manuscripts with musical notation available online, ONRP has begun this process using a tenth-century manuscript, called the *Hartker Antiphoner*, as a prototype. The manuscript comes from the St. Gallen Stiftsbibliothek in Switzerland and has an early chant notation that uses neumes without staff lines. Notated by hand, this Gregorian chant notation poses many research challenges. Gregorian chant shows in its essence the relationship between word and music. As a regular

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 275–291.

part of the liturgy, plainchant sets biblical texts to music; given the importance of the texts, they are understandably set very carefully and sensitively. For example, even if a melody is used for two different texts, the notation of the melody changes to reflect the different structure and meaning. With this level of sensitivity, how can we capture the variety of over 300 neume signs and their combinations that we find in the *Hartker Antiphoner*? For this technical issue, we use the *Music Encoding Initiative* (MEI), which is a markup language that has become the standard for the discipline of music. Each sign in a digital image is described within a consistent hierarchy of attributes. MEI is adaptable to every kind of notation because the attributes can be determined freely. Thus, MEI has been used for notations ranging from tablature, to mensural notation, to *Hufnagelnotation* (Morent, Tübingen). The following article describes the process of adaptation of MEI for neume notation without staff lines. In order to explain clearly where the challenges are, the characteristics of the neume notation are briefly described, as well as how the *classifier* (a kind of digital neume dictionary) works.

1 Das *Optical Neume Recognition Project* (ONRP) und die MEI Kodierung

Das *Optical Neume Recognition Project* (ONRP) ist ein Zweig von *Cantus Ultimus,* einem Forschungsteam, das zu dem Verbund verschiedener Forschungsprojekte SIMSSA (*Single Interface for Score Searching and Analysis*) gehört, die unter der Leitung von Ichiro Fujinaga stehen und an der McGill University in Montreal angesiedelt sind. *Cantus Ultimus* beabsichtigt, die optische Wiedererkennungstechnologie für Bilddateien mit Musiknotation weiterzuentwickeln. Das *Optical Neume Recognition Project* (ONRP) ist dasjenige Teilprojekt, das insbesondere mit früher linienloser Notation in Neumen aus Handschriften etwa des 10. bis 12. Jahrhunderts befasst ist.

Im Rahmen der Forschungsarbeiten von *Cantus Ultimus* sollen Tools erstellt werden, die helfen, nach musikalischer Information in Bilddateien zu suchen – vergleichbar zur optischen Textwiedererkennung, wie es beispielsweise bei *Google Books* geschieht. Mit *Cantus Ultimus* wurde bereits jetzt eine Oberfläche erstellt, genannt *Cantus Ultimus viewer,* auf der Bildseiten von mittelalterlichen Handschriften zu sehen sind, kombiniert mit Informationen zu den Inhalten auf diesen Handschriftenseiten. Diese Informationen stammen aus der bereits existierenden Datenbank *Cantus Manuscript Database*. Für diese Datenbank wurden seit den 1980er Jahren an der *Catholic University of America* unter Leitung von Ruth Steiner mittelalterliche Choralhandschriften katalogisiert. Heute sind mehr als 160 mittelalterliche Codices mit lateinischem, einstimmigen Repertoire des Gregorianischen Chorals in dieser Datenbank enthalten. In den letzten Jahren wurden von einem Teil der Handschriften von wissenschaftlichen

Mitarbeitern der Text wie auch die Gesänge vollständig transkribiert und somit die Datenmenge der *Cantus Manuscript Database* entscheidend erweitert, da hierdurch für die Forschung wichtige neue Suchoptionen ermöglicht worden sind: Auf dieser Grundlage bietet *Cantus Ultimus* heute neben der Abbildung der einzelnen Handschriftenseiten (als Scan), Informationen zum Inhalt und verschiedene Suchmöglichkeiten. Die Suchergebnisse können sogleich in allen Handschriftenseiten nachgeschlagen werden, was den Umgang mit dem Codex enorm vereinfacht.

Eine besondere Herausforderung für *Cantus Ultimus* ist die große Unterschiedlichkeit der musikalischen Notationsstile in den Gregorianikhandschriften. Denn die Handschriften stammen aus ganz verschiedenen Regionen Europas und aus einem Zeitraum, der mehr als 600 Jahre umfasst. Trotzdem lassen sich die Handschriften dieser verschiedenen Notationsstile in zwei große Gruppen aufteilen: in diejenigen Handschriften mit Notation auf Linien, die Informationen zur absoluten Tonhöhe der Melodietöne bieten (*diastematische Notation*), und diejenigen Codices ohne Notation auf Linien, die keine absolute Tonhöhe der Einzeltöne anzeigen (*adiastematische Notation*), sondern pro Silbe die Anzahl der Töne, ihre Relation zueinander (hoch, tief, gleich hoch), sowie rhythmisch-agogische Aspekte beschreiben. Jede dieser Notationen erfordert verschiedene Forschungsansätze, so dass eine Gruppe von *Cantus Ultimus* mit Handschriften mit Liniennotation arbeitet, wohingegen das *Optical Neume Recognition Project* den Fokus auf die Handschriften ohne Liniennotation gelegt hat.

Das Projekt startete mit einer Handschrift des 10. Jahrhunderts in zwei Bänden, dem berühmten *Antiphonar Hartker,* Stiftsbibliothek St. Gallen, CH-SGs 390 und 391, das um 1000 entstanden ist und in Verbindung gebracht wird mit dem St. Galler Mönch Hartker. Diese Handschrift wurde gewählt, da sie eines der frühesten vollständig erhaltenen Antiphonare darstellt und bereits viele Studien zum Codex, beispielsweise hinsichtlich der verschiedenen Schreiber der Notation, erstellt worden sind – Studien, die mit Hilfe von ONRP verifiziert und ergänzt werden sollen (Pouderoijen und de Loos 2009). Die Notation des *Antiphonars Hartker* repräsentiert ein frühes Stadium der Notation mit einer hohen Vielfalt und einem hohen Variantenreichtum der Notationszeichen.

Im *Antiphonar Hartker* sind die Neumen als aufrechte Strichnotation notiert, eine sogenannte *Akzentnotation,* die an einer gedachten horizontalen Linie oberhalb der jeweiligen Textzeile ausgerichtet werden. Dabei erscheinen die Neumenzeichen ohne Linien in das freie Feld *(in campo aperto)* über den Textzeilen. Die Neumen bieten Informationen über die relative Tonhöhe der Melodie, wie weiter unten beschrieben werden wird, und sie bietet außerdem eine große Menge an weiteren Informationen über die Aufführungsweise, und zwar hinsichtlich des Tempos, der Tondauer und Dynamik – dies jeweils mit engstem Bezug zum Text. In unserer heutigen Notation sind wir gewohnt, dass Informationen zur Tonhöhe und Tondauer getrennt gegeben werden von denjenigen über die Art der Aufführung, also zur Art und Weise, *wie*

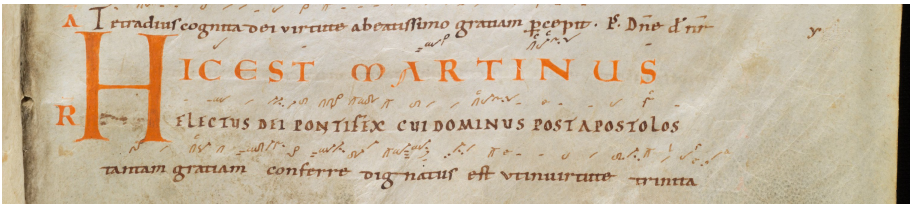


Abbildung 1: Neumen des Responsoriums *Hic est Martinus* vom Fest des Heiligen Martins, *Antiphonar Hartker* CH-SGs 391, um 1000, St. Gallen, Stiftsbibliothek, Cod. Sang. 391, pag. 141.

ein Ton ausgeführt werden soll. In der frühen Neumennotation gibt es aber kaum Zeichen, die neben Angaben zur Aufführungsweise nicht auch solche zur Tondauer und Tonhöhe machen. Wenn man nun versucht, mit Hilfe von Computertechnik die Notation zu erfassen, muss dieser Andersartigkeit des Konzeptes der Notation Rechnung getragen werden. Es müssen zunächst möglichst viele Informationen, die in der Notation in Neumen enthalten ist, beschrieben werden, damit sie im Anschluss in verschiedenster Kombination von der Computertechnik erkannt werden können.

Über 300 verschiedene Neumen – verstanden in diesem Fall als einzelnes Zeichen wie auch als Kombination von Zeichen – wurden im *Antiphonar Hartker* gefunden. Schätzungsweise 150.000 verschiedene Neumenzeichen sind in den zwei Bänden enthalten. (Berechnet wird dies überschlagsweise mit einer durchschnittlichen Anzahl von 20 Neumen pro Zeile bei durchschnittlich 15 Zeilen pro Seite und insgesamt 502 Seiten.) Sobald man nach bestimmten gleichen Neumenkonstellationen in vielen hundert Bildseiten suchen kann, d. h. nach Mustern der Neumierung, wird sich die wissenschaftliche Erforschung der Neumennotation immens verändern. Weitere Muster werden erkannt werden, solche hinsichtlich der Verwendung der Musiknotation, bei der Herstellung der Handschrift, solche hinsichtlich der Verankerung der Gesänge in der Liturgie, in der dieser Gesänge erklungen sind, sowie hinsichtlich des engen Wort-Ton-Verhältnisses der Gesänge.

2 Music Encoding Initiative (MEI) – Zur Kodierung der Notationszeichen

Eine der besten Möglichkeiten zur Beschreibung der *Music Encoding Initiative* (MEI) ist, diese mit einer menschlichen Sprache zu vergleichen. Sprachen sind Kodierungssysteme, die auf Regeln beruhen, die von denjenigen eingehalten werden, die sich der Sprache bedienen. Wir nennen diese Regeln die Grammatik der Sprache. Eine Grammatik benötigt sowohl Freiräume als auch semantische Eindeutigkeit, und sie muss auch flexibel genug sein, um neue Vokabeln aufzunehmen. Die Regeln (die Gram-

matik) von MEI betreffen die Art und Weise, wie die Beschreibung der musikalischen Notation aufgebaut ist und auch welche Arten von Kennzeichen (Attributen) in diesen Beschreibungen erlaubt sind. MEI hat sich bewährt bei der Vermittlung zwischen dem Verständnis von Notation durch den Musikforscher und den Möglichkeiten, diese Vorstellungen mit Hilfe von Computertechnologie wiederzugeben. Und wie bei einer Sprache benötigt man Zeit und Einsatz, diese zu lernen und aktiv zu beherrschen. In unserem Fall bedeutet dies das Überdenken der Bedeutung von Neumennotation und ihrer Zeichen und des intuitiven Verständnisses der Einzelzeichen, bis man schließlich zu demjenigen gelangt, was durch den Computer dargestellt werden kann. Sobald diese Phase durchschritten ist, sind die Vorteile der Verwendung von MEI bestechend. Dahinter steht eine weltweit organisierte Gemeinschaft von Bibliothekaren, Historikern, Musiktheoretikern, Musikwissenschaftlern und Computerspezialisten. MEI beruht auf den Prinzipien von XML (*Extensible Mark-up Language*). Die MEI-Community hat sich den Idealen Transparenz und freie Verfügbarkeit verschrieben, bekannt unter dem Schlagwort *open source*. Die Grammatik von MEI basiert auf einem strikten System von Hierarchien, so wie dies bei allen Elementen der XML-Familie der Fall ist. Wenn man erneut den Vergleich zur Sprache bemüht, ist der Prozess des MEI-Codes ähnlich der Struktur eines Satzes: Es gibt Elemente, die wesentliche und unverzichtbare Elemente sind wie ein Subjekt und ein Prädikat im Satz. Andere Elemente, wie ein Nebensatz beispielsweise, sind optional. MEI ist erweiterbar, da in die hierarchische Struktur weitere Elemente mit aufgenommen werden können – so beispielsweise Zeichnungen. Alle Dinge, die mit einer Serie von Elementen in einer hierarchischen Struktur beschreibbar sind, können mit XML kodifiziert werden.

Die MEI Kodierung erfordert das Einhalten einer standardisierten Hierarchie. Sie macht aber keine Vorgaben hinsichtlich der Art der Merkmale der Notationszeichen, die enthalten sein müssen. Auf diese Weise kann sie adaptiert werden auf verschiedenste Notationsarten, wie beispielsweise Tabulaturschrift, Mensuralnotation und sogar auf Neumennotation. Für die Mitarbeiter von *Cantus Ultimus* ist entscheidend, dass wir mit offenen und transparenten Methoden arbeiten, damit die MEI der linienlosen Neumennotation möglichst von vielen genutzt oder für andere Notationstypen angewendet werden kann.

Mehrere Jahre zuvor hat Stefan Morent (Tübingen) erstmals MEI angewendet bei Notation mit Neumen auf Linien, namentlich Hufnagelnotation, in seinem Projekt *Digital Music Edition* (DiMusEd). Er hat dabei Gesänge von Hildegard von Bingen in MEI erfasst. Das Beispiel zeigt einzelne Teile dieser MEI Kodierung. Hierbei wirkt die Textsilbe als oberste Bezugsgröße der Struktur. Auf der nächsten Ebene werden die Notationszeichen pro Silbe angegeben. Im oberen Teil des Beispiels ist ein Porrectus (`uneume name`) angegeben mit den Tönen e, d, e (notiert zu den Attributen `oct` und `pname`), die zur Textsilbe »O« gesungen werden. »O« ist die erste Silbe des Gesangs und wird in der Handschrift notiert als rote (angegeben als `<rend color="red">`)

Initiale (angegeben als `<syl n="initial">`). Die folgende Textsilbe im unteren Teil des Beispiels zur Silbe »splen« trägt zwei Neumen: eine Clivis (Tonhöhen g-e) und einen Pes (Tonhöhen d-e).[1]

```
<layer>
 <syllable>
  <syl n="initial">
   <rend color="red">0</rend>
  </syl>
  <uneume name="porrectus">
   <note oct="3" pname="e"/>
   <note oct="3" pname="d"/>
   <note oct="3" pname="e"/>
  </uneume>
 </syllable>
 <syllable>
  <syl>splen_</syl>
  <uneume name="clivis">
   <note oct="3" pname="g"/>
   <note oct="3" pname="e"/>
  </uneume>
  <uneume name="pes">
   <note oct="3" pname="d"/>
   <note oct="3" pname="e"/>
  </uneume>
 </syllable>
</layer>
```

Beispiel: Hufnagelnotation; Neumen auf Linien in MEI

Die hierarchische Struktur von XML mag auf den ersten Blick schwerfällig wirken, und man fragt sich vielleicht, warum nicht auf kommerzielle Musiksoftware wie beispielsweise dem Schreibprogramm *Finale* zurückgegriffen worden ist oder auf frei zugängliche Schrifttypen wie *Volpiano*. MEI hat mehrere Vorteile: Der erste Vorteil von MEI besteht darin, dass es eine *open source*-Sprache ist. Alle Elemente, die notwendig sind, um zu verstehen, wie MEI funktioniert, sind frei zugänglich. Alle Möglichkeiten zur Weiterentwicklung, zur Adaptierung auf die eigenen Bedürfnisse, stehen frei zur Verfügung. Das Prinzip *open source* ist nicht nur eine politische Überzeugung, sondern auch eine technische Errungenschaft. MEI als *open source*-Technologie weltweit als Standard zu verankern, bedeutet, dass es nicht zu kommerziellen Zwecken missbraucht werden kann. Jede Notationsart wird einmal mit MEI erfasst werden können. Der zweite Vorteil von MEI ist, dass es das Potential hat, einmal zur Standardsprache der Repräsentierung von Musiknotation in einer computergestützten Technologie zu werden. Heute ist es wichtig, dass wir Standards international vereinbaren. Wir benötigen zukünftig eine Sprache, bei der nicht individuelle Methoden vorherrschen, die nur für bestimmte Forschungsprojekte und -vorhaben anwendbar sind, sondern solche, die leicht kompatibel und erweiterbar sind, so dass sie auch von anderen Forschungsprojekten genutzt werden können. Täglich wächst die Anzahl digitaler

[1] Aus Platzgründen wird hier darauf verzichtet, das Wortende ebenfalls zu beschreiben.

Bilddateien von (mittelalterlichen) Handschriften im Internet. Es ist daher wichtig, dass Standards hinsichtlich der Methodik, wie die musikalische Information dieser Bilder erfasst werden soll, eingeführt werden. Einem einzelnen Forscher ist kaum noch möglich, die immens große Menge an digitalen Abbildungen zu überblicken; mit MEI als Standardkodifizierung wird es möglich sein, mit einer großen Menge an Musikdaten in der Zukunft effektiv zu arbeiten. Vielleicht können einmal tausende von Gesängen verglichen werden sowie vollständige Handschrifteninhalte analysiert werden auf der Suche nach wiederkehrenden Strukturen. In jedem Fall ist die Flexibilität der Kodierung ein wesentlicher Faktor. Früher voneinander unabhängige Forschungsprojekte, die zu verschiedenen Themenbereichen der Choralforschung arbeiten, werden zukünftig profitieren können durch den Austausch, der durch MEI ermöglicht wird.

3 Die Notation in Neumen und ihre Herausforderung

Das *Antiphonar Hartker* der Stiftsbibliothek ist eine der frühesten vollständig mit Musiknotation versehenen Handschriften, die uns überliefert ist. Das in diesem Codex enthaltene Musikrepertoire gehört zum Gregorianischen Choral, einem einstimmigen liturgischen Gesang in lateinischer Sprache. Dieses Repertoire wurde, so der heutige Kenntnisstand, zunächst mündlich überliefert. Erste schriftliche Zeugnisse gehen auf das achte Jahrhundert zurück. Das *Antiphonar Hartker* aus dem Jahr 1000 stammt aus der Blütezeit der adiastematischen Notation in St. Galler Neumen. Ohne an dieser Stelle in größere Details eingehen zu können, sei aber doch erwähnt, dass diese linienlose Notation in der Regel keine genauen Intervalle der einzelnen Töne der Melodie angibt. Die Melodie war, wenigstens zum größten Teil, noch hinsichtlich ihres Tonhöhenverlaufs bekannt. Der Reichtum der Notation zeigt sich in Aspekten wie den Folgenden: Die Neumen geben Hinweise hinsichtlich der Tondauer und auch der agogischen Gestaltung der Melodien. Auch gibt es manche Neumenzeichen, die die Aussprache der Vokale und Konsonanten anzeigen, sowie Zusatzbuchstaben bei den Neumen, die beispielsweise Angaben zur Lautstärke sind. Insgesamt zeigt sich durch den Informationsreichtum der Neumenotation sowie durch den Melodieverlauf selbst, dass der Text differenziert betont und vorgetragen werden soll. In der Phase der mündlichen Überlieferung bleibt es nicht bei einer einfachen gesungenen Textrezitation, sondern es entwickelt sich eine aufwendige Melodiegestalt und die einzelnen in den Handschriften notierten Gesänge spiegeln eine bestimmte Textinterpretation wieder.

Umfangreiche Melodierestitutionen der Gesänge des Messrepertoires sind nach dem zweiten Vatikanischen Konzil erstellt worden durch den Vergleich der frühen linienlo-

Punctum		tief
Gravis		tief, Abstieg um mindestens eine Terz
Virga		hoch, aufsteigend oder absteigend
Clivis		hoch-tief
Pes		tief-hoch
Porrectus		hoch-tief-hoch
Torculus		tief-hoch-tief

Tabelle 1: Einige Grundzeichen der St. Galler Neumennotation

sen Handschriften mit solchen späterer Jahrhunderte mit Liniennotation.[2] So können die Gesänge rekonstruiert werden hinsichtlich ihres melodischen Tonhöhenverlaufs und hinsichtlich ihrer rhythmisch-agogischen Gestaltung und der liturgischen Praxis in einer Fassung zur Verfügung gestellt werden, die eine größtmögliche Annäherung zur Notation in frühen Neumen darstellt.

Einige Grundzeichen der St. Galler Neumennotation seien in Tabelle 1 dargestellt.

Wie sogleich erkannt werden kann, geben die Grundzeichen der Notation die Anzahl der Töne an sowie die Melodierichtung oder eine melodische Kontur.

Die Neumenzeichen können modifiziert werden, indem sie gedreht werden (so beispielsweise ein *Torculus:* ), in dem sie vergrößert werden (so beispielsweise der liqueszierende *Climacus,* genannt *Ancus:* ) oder indem ihnen etwas hinzugefügt wird (so beispielsweise die liqueszierende *Virga:*  oder eine *Virga mit Episem:* ). Diese Modifikationen des Grundzeichens haben, wie in vielen Studien von Musikwissenschaftlergenerationen erforscht worden ist, Auswirkungen auf die Anzahl der Töne (Als *Liqueszenz* wird beispielsweise bezeichnet, wenn bei bestimmten Konsonantenkonstellationen zwei aufeinanderfolgende Silben sehr eng miteinander verbunden werden, wobei zusätzliche Zwischentöne entstehen können.) und auf die rhythmische Ausführung der Töne. Zusätzlich können sogar vereinzelt Informationen zu einzelnen Intervallen gegeben werden (Engels 1998; Engels 2006).

Einige Zeichen sind in ihrer genauen Bedeutung noch nicht erschlossen. So stehen die Forscher und Musiker beispielsweise fragend vor dem Notationszeichen *Oriscus,* das isoliert () oder mit vielen anderen Zeichen kombiniert erscheint, und diskutieren auch über das Phänomen der *Liqueszenz.*

[2] Vergleiche die Arbeiten der Melodierestitutionsgruppe der *Gesellschaft für Studien des Gregorianischen Chorals* (AISCGre), die durch den Solesmer Mönch Dom Eugéne Cardine angestoßen worden sind, sowie die Restitutionsarbeiten von Alberto Turco.

Die Software, die für *Cantus Ultimus* entwickelt wird, kann schlussendlich nicht die offenen Fragen beantworten. Aber sie kann helfen, bestimmte Muster und Strukturen in der Notation zu erkennen. Beispielsweise kennt man bereits Neumen, die zum Teil verbunden und zum Teil unverbunden notiert werden, so der *Porrectus* und die *Clivis* ( sowie ) und fragt sich, ob dies eine Gewohnheit eines bestimmten Schreibers ist oder eine bestimmte (rhythmische?) Qualität hat. Auch sind die Regeln der sogenannten *Neumentrennung* bekannt in Melismen über einer Textsilbe, nach der in auf- und absteigenden Melodiepositionen sowie an den höchsten Stellen der Melodie die getrennte Schreibweise von Notationselementen eine Dehnung der jeweils letzten Note vor der Trennung zum nächsten Zeichen ausdrückt. Aber hält sich jede Handschrift der St. Galler Handschriftenfamilie an diese Regel? Eine größere Feldstudie wäre hier wünschenswert.

Eine *Big Data*-Suchmaschine kann helfen, gleichbleibende Muster wiederzuerkennen. Wie hilfreich ist dies in einem Repertoire, das mündlich überliefert worden ist und in dem viele gleichbleibende Elemente vorhanden sind, die dem Sänger halfen, die vielen Gesänge mit den vielen kleinen Melodiewendungen zu memorieren!

Beispielsweise findet sich im Repertoire eine stabile Verwendung bestimmter Spezialneumen: Der *Intonationstorculus,* notiert mit einer ausladend vergrößerten Torculusneume, steht für drei Töne, wobei die erste Note zur zweiten Note eine aufsteigende kleine Terz oder Quart ist und die zweite zur dritten Note stets ein Halbton. Die Neume kann in anderen adiastematischen Notationsfamilien, wie der lothringischen adiastematischen Neumenotation, sehr gut erkannt werden. In der St. Galler Familie wird die Neume nicht immer notiert, d. h. häufig findet sich die übliche Torculusneume ohne Modifikation. – Warum ist dies so? Gibt es bestimmte Schreiber, die davon ausgehen, dass der *Intonationstorculus* erkannt wird, auch ohne, dass er eigens durch Veränderung des Zeichens angezeigt werden muss? – In der heutigen Aufführungspraxis wird, sofern diese sich an der Neumennotation orientiert, dieses Neumenzeichen durch eine starke Dehnung der zweiten und dritten Note ausgeführt. Denn stets steht der Wortakzent auf der übernächsten Silbe, und die Note über dem Wortakzent ist einen Ton höher als die mittlere Torculusnote. Dieser *Intonationstorculus* schafft auf diese Weise einen rhythmischen Stau, indem er vor der entscheidenden Silbe mit Wortakzent eine Dehnung erzeugt, gerade, um das gesamte Wort hervorzuheben. Häufig findet man diese Spezialneume bei wichtigen sinntragenden Worten des Gesangtextes. – Zusammenfassend lässt sich sagen, dass diese Neume für eine bestimmte Verteilung der betonten und unbetonten Silben auf eine bestimmte Anzahl von Melodietönen steht, dass eine bestimmte Tonfolge (Quart oder kl. Terz + kl. Sekunde) durch die Neume angegeben wird und eine bestimmte rhythmische Gestaltung vorliegt und diese mit einer bestimmten Gestaltung der gesamten Textstelle einhergeht. Dies sind schon viele Informationen für einen Sänger, der den Gesang auswendig memoriert. – Aufgrund der klaren Intervallstruktur kann der *Intonationstorculus* nur an bestimmten Stellen

in der Tonskala auftreten. Wie häufig er in welchen Modi vorkommt, wäre in diesem Fall nicht ein Erkenntnisgewinn zum Wesen des *Intonationstorculus.* Aber die Häufigkeit solcher und ähnlicher musikalisch wiedererkennbarer und gleichbleibender Strukturen könnte die Stilistik bestimmter Repertoireschichten beleuchten. Wurden *Intonationstorculi* mit gleicher Häufigkeit in den Neuschöpfungen, den Gesängen in Heiligenoffizien, des 11. und 12. Jahrhunderts verwendet?

Notation ohne Linien in Neumenschrift eröffnet bereits viele Fragen, die teilweise bis heute nicht geklärt sind. Es gibt aber sogar ein Repertoire, der altspanische Choral, notiert in sogenannten *mozarabischen* Neumen *in campo aperto,* dessen Melodien wir zum größten Teil nicht mehr rekonstruieren können. Denn diese Gesänge sind in keinen Quellen mit Liniennotation enthalten. So fehlt für die Rekonstruktion der Melodien, d. h. der genauen Intervallfolgen, eine verlässliche Quelle.

Das altspanische Choralrepertoire ist in fünf Handschriften enthalten, wovon die frühesten aus dem 10. Jahrhundert stammen (Hornby und Maloy 2013). Nur für etwa ein Dutzend Gesänge ist bekannt, welche Melodien die Neumen ausdrücken. Dieser besondere Tatbestand wie auch das Wissen um die reiche Tradition der mündlichen Überlieferung haben Emma Hornby und Rebecca Maloy motiviert, diese Notation zu untersuchen. In ihren Forschungsprojekten (*Compositional Planning, Musical Grammar and Theology in Old Hispanic Chant,* 1/11/09 bis 1/05/11 und *Understanding Old Hispanic chant manuscripts and melodies,* 1/07/16 bis 1/07/17) erforschten Sie und ihr Team den Zeichenschatz der Neumenschrift sowie die Häufigkeit der Verwendung bestimmter Neumengruppen und deren Position innerhalb des Satzes. Sie sind somit den Notationsgesetzmäßigkeiten äußerst nahe auf die Spur gekommen und haben die Grammatik der altspanischen Musiknotation umfangreich beschrieben.

Obwohl aus genannten Gründen nicht möglich ist, Tonhöhen den einzelnen Neumenzeichen der altspanischen Notation zuzuordnen, konnte aber trotzdem jedes Einzelzeichen hinsichtlich der melodischen Kontur, für das es steht, beschrieben werden. Die Buchstaben N (= neutral), H (= high), L (= lower), S (= same) geben die relative Position der Note in der melodischen Kontur an und bildete den Ausgangspunkt für ihre hoch differenzierte Beschreibung der einzelnen Notationszeichen. Wurden einmal alle mozarabischen Neumenzeichen auf diese Weise erfasst, erschlossen sich durch den Vergleich derjenigen Neumen, die für dieselbe melodische Kontur stehen, die aber verschieden graphisch notiert sind, neue Erkenntnisse. So verhalf die Mustererkennung mithilfe eines zweiten Codes und die Sortierhilfe durch den Computer zum Erkenntnisgewinn. Im *Optical Neume Recognition Project* haben wir erkannt, dass diese Beschreibung der Einzelzeichen mit den Repräsentanten N, H, L und S für alle Notationen mit relativer Tonhöhenangabe genutzt werden kann. Natürlich ist es nur ein bestimmtes Merkmal, das den Zeichen dieser Notationen gemein ist. Ein anderes gemeinsames Element wäre für die Akzentnotation, zu der auch die St. Galler Neumennotation gehört, ihre diagonale Schreibrichtung (Corbin 1977, s. Abb. 2). Doch

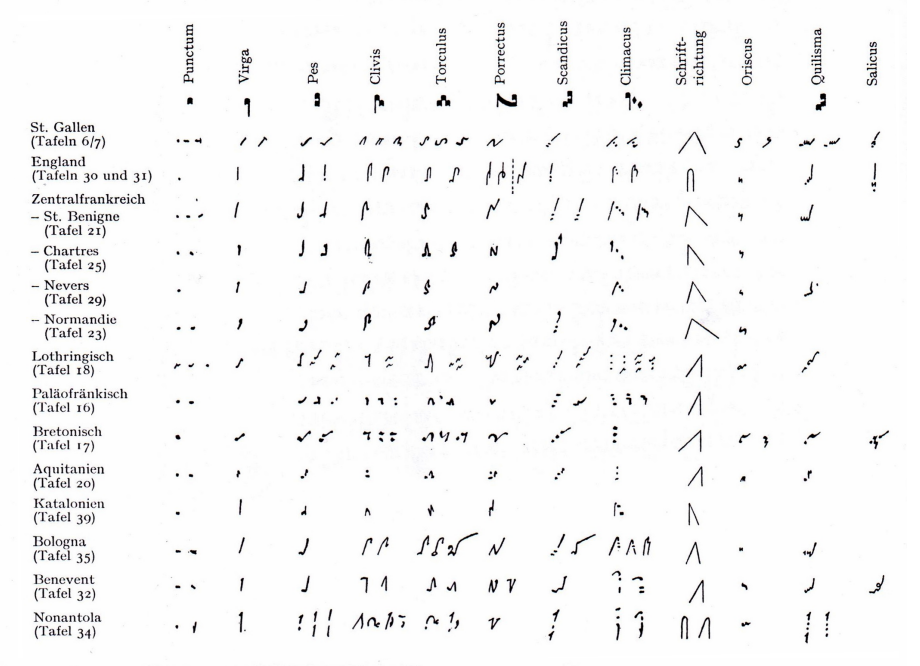


Abbildung 2: Neumentabelle von Corbin (1977).

die Notierung der Neumenzeichen in einer Abfolge von N (neutral, or unknown), H (higher), L (lower), S (same, or unison), A (same, or higher) und U (same, or lower) bietet den Vorteil der Vergleichbarkeit mit Notationen auf Linien. Aufgrund dessen haben wir die 114 einzelne Komponenten, die wir im *Antiphonar Hartker* gefunden haben, erfasst in einer Abfolge dieser Buchstaben. Ein *Punctum* ( ) steht beispielsweise für eine Note mit undefinierter (neutral) Tonhöhenbeziehung zum Folgeton und wird daher mit »1-N« angegeben. Zusätzlich geben Kleinbuchstaben Charakteristika der Einzeltöne an: w (wavy), b (curved anticlockwise), c (curved clockwise), a (angled), e (episema), f (flat), j (jagged), l (liquescent), x (extended), y (diagonal right up), k (diagonal right down), q2 (quilisma 2 curves), q3 (quilisma 3 curves). Ein *Tractulus* ( ) wird daher mit »1-Nf« beschrieben, ein *Tractulus mit Episem* ( ) mit »1-Nfe«. Die *Gravis* ( ) erhält die Folge »1-Nfk« und die *Stropha* ( ) die Abfolge »1-Nc«. Schon die frühen Neumentabellen ab dem 12. Jahrhundert haben einzelnen Zeichengruppen als eine Neume benannt (Floros 1970, S. 184-207). So ist (in diesem Fall eine *Virga* mit einem Episem) nicht eine *Virga mit zwei Puncta,* sondern erhielt den Namen *Climacus.*

Gemeint sind drei absteigende Töne. Um in unserer Kodierung die Trennung der Einzelzeichen innerhalb der Neume anzugeben, nutzen wir den Buchstaben g (gapped). Zusätzlich gibt es Neumen, nach denen die Melodie tiefer oder höher weitergeht; dies geben wir an mit d (down, afterwards lower) oder u (up, afterwards higher). Zu den Neumenzeichen notieren manche Schreiber in den Handschriften auch Buchstaben, die häufig für die Aufführung der Noten relevant sind, so beispielsweise c für *celeriter* (schnell). Solche Zusatzbuchstaben werden ebenfalls angegeben, im Fall von c mit »p:c«.

4 Der *Classifier* – Ein digitales Neumen-Wörterbuch

Auch wenn wir mit MEI die einzelnen Neumen des *Hartker Antiphonars* akkurat wiedergegeben haben, muss immer noch die MEI Kodierung verbunden werden mit dem betreffenden Bildausschnitt im digitalen Scan der Handschriftenseite. Es bleibt, um es auf den Punkt zu bringen, die Frage nach der Verbindung der beiden Wege, mittels derer der Computer Input erhalten kann: einerseits durch Bildverarbeitung und andererseits über den Code.

Im Fall der Handschrift Hartker wurde die Bildverarbeitung bewerkstelligt als Teil eines Systems genannt *Rodan,* welches von Andrew Hankinson im Musiktechnologie-Labor von SIMSSA an der McGill University in Montreal entwickelt worden ist. Jedes Bild ist automatisch binarisiert, was bedeutet, dass in mehreren Bildbearbeitungsprozessen der zunächst farbige Scan der Handschriftenseite in einen schwarz-weiß Scan verwandelt wird. In einem weiteren Schritt müssen schwarze Pixeln, die am Rand durch den Bilduntergrund der Pergamenthandschrift entstanden sind und vom Computer als Notation missdeutet werden könnten, erkannt und gelöscht werden. Hierauf liest eine Software, die auf der Software *Gamera* basiert und OMR-Algorithmen verwendet, auf der Handschriftenseite gleichbleibende Abfolgen von schwarzen und weißen Pixeln und sortiert diese nach ihrer Ähnlichkeit. Die Einzelzeichen werden sortiert gemäß einer Tabelle von Notationszeichen, die wir wie eine Art Neumen-Wörterbuch für die Software verfasst haben und *Classifier* nennen. Mit diesem Wörterbuch kann die Software die Funde jeweils vergleichen. In dem Moment, wo die noch nicht identifizierte Abfolge von schwarzen und weißen Pixeln einem Eintrag im Wörterbuch zugeordnet worden ist und damit einen Namen erhält, ist es möglich, auch den umgekehrten Weg zu gehen und dem einen bestimmten Ausschnitt im digitalen Bild einen MEI-Code zuzuteilen. Dies ist der Verbindungspunkt zwischen Bild und Kodifizierung.

In diesem Prozess lernt der *Classifier* durch die Expertise der Musikwissenschaftler, und umgekehrt profitiert der Musikwissenschaftler von der Schnelligkeit der Datenerfassung durch den Computer. Zunächst benötigt der Computer den Musikwis-

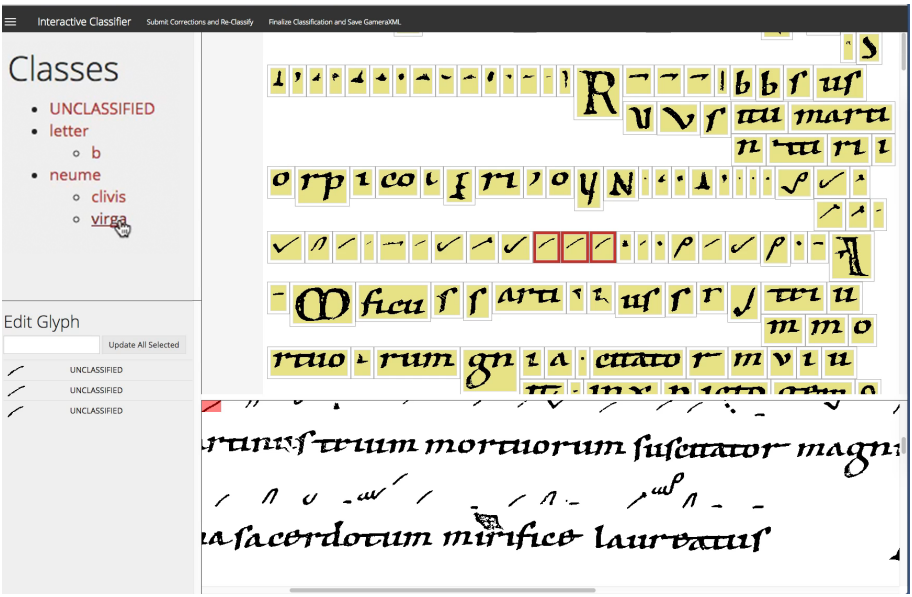


Abbildung 3: Der Musikwissenschaftler markiert drei Neumen im Feld rechts und benennt sie als *Virga,* s. links in dem Beispiel.

senschaftler, um für jeden Eintrag im Neumen-Wörterbuch ein Beispiel zu erhalten. Dieselbe Handschriftenseite wird dann erneut mit der Software durchsucht. Im zweiten Durchlauf wird die Trefferquote schon weit höher sein, da das Neumen-Wörterbuch mehr und mehr Eintragungen erhält in diesem Prozess.

Dem Musikwissenschaftler kommt in diesem Prozess eine wichtige Rolle zu: Er muss interpretieren. Die reine Sortierung nach Mustern weißer und schwarzer Pixel reicht nicht aus. Der Forscher erkennt, ob das Zeichen gegebenenfalls durch den Prozess der Binarisierung ursprünglich anders ausgesehen hat, ob ihm gegebenenfalls ein Teilelement fehlen könnte. In einem solchen Fall muss er im farbigen Digitalisat nachsehen, wie das Notationszeichen dort aussieht und kann dann die richtige Zuweisung im Neumen-Wörterbuch vornehmen.

In diesem Prozess können große Mengen von Neumen auf einen Blick gesehen werden. Bestimmte Gewohnheiten der verschiedenen Schreiber in einer Handschrift können so mit Hilfe des *Classifiers* beobachtet werden. In einem späteren Stadium, wenn weitere Neumenhandschriften mit MEI erfasst sind, können wohlmöglich auch Schreibtraditionen innerhalb beispielsweise der St. Galler Handschriftenfamilie untersucht werden.

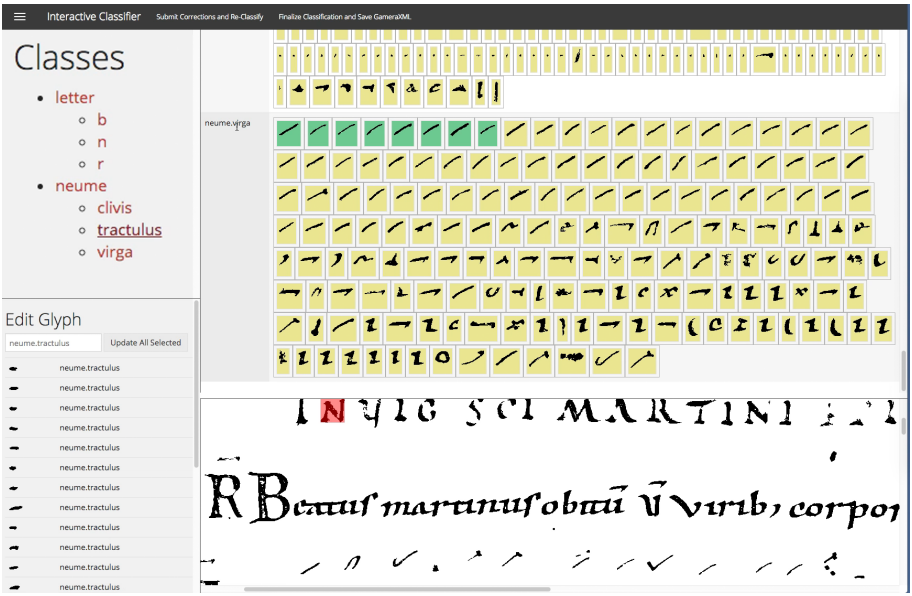


Abbildung 4: Der *Classifier* sortiert die Zeichen nach Grad ihrer Ähnlichkeit zu derjenigen *Virga,* die der Musikwissenschaftler als Prototyp zuvor markiert hat.

Ein Vorteil dieses Vorgehens ist, dass die Neumenzeichen der Handschrift nicht zunächst standardisiert werden müssen. Der Computer erfasst alle Zeichen auf der Handschriftenseite, so wie sie dort zu finden sind. Dies ermöglicht den Vergleich desselben Notationszeichens in verschiedensten graphischen Modifikationen.

Beim Prozess der Binarisierung geht naturgemäß viel Information verloren. Deshalb wird der *Classifier* schlussendlich allein so gut sein, so gut händisch Falschzuweisungen der Software gefunden worden sind. Hier ist der Musikwissenschaftler gefragt, der interpretieren und erkennen kann.

Das Wörterbuch besteht momentan aus 114 unverbundene einzelne Komponenten der Notationszeichen. Als nächster Schritt soll dem *Classifier* gezeigt werden, welche Kombinationen von Zeichen über einer Textsilbe des Gesangs möglich sind, d. h. welche »Neumen« es gibt. In der Gregorianikforschung meint der Begriff »Neume« (*neuma,* griech. - der Wink, die Geste) im engeren Sinn alle Musikzeichen als Gesamtheit, die über einer Silbe stehen. Beide Informationen zusammen – jede einzelne Komponente wie auch alle möglichen Kombinationen über einer Textsilbe – lassen schlussendlich den *Classifier* alle Muster der Gruppierung der einzelnen musikalischen Zeichen erkennen.

5 Vom Öffnen der *Büchse der Pandora – Big Data* der kleinen und kleinsten Zeichen der Notation in Neumen

Nach der griechischen Mythologie entwichen aus der *Büchse der Pandora,* als diese von ihr geöffnet worden war, alle Übel, Laster und Mühen in die Welt. Die erste Frau, Pandora, erhielt die Büchse vom Göttervater Zeus mit dem strengen Hinweis, diese unbedingt nicht zu öffnen. Doch Pandora öffnete die Büchse. – Aus Neugier? Getrieben von der Sehnsucht, vielleicht Gier, im Gefäß etwas besonders Schönes oder Edles zu finden? – Nichts dergleichen sollte sie finden. Vielmehr entwichen sogleich Krankheiten, Arbeit, Negatives, sogar Tod dem Gefäß und erfassten die gesamte Schöpfung. Und schließlich entwich auch die Hoffnung dem sonderbaren Gefäß. Doch ist die Hoffnung ein Übel? Für manche Menschen bleibt kaum mehr als Hoffnung, um die Gegenwart zu durchdauern. Das Prinzip Hoffnung ist für viele Menschen Ausdruck des Glaubens bzw. Glaube schlechthin: »Die Schöpfung ist der Vergänglichkeit unterworfen, nicht aus eigenem Willen, sondern durch den, der sie unterworfen hat; aber zugleich gab er ihr Hoffnung.« (Römerbrief 8,20). Anders äußert sich Nietzsche, für den Hoffnung erst der Anfang allen Übels ist, wenn er direkt Bezug nimmt auf Pandora: »Zeus wollte nämlich, dass der Mensch, auch noch so sehr durch die anderen Übel gequält, doch das Leben nicht wegwerfe, sondern fortfahre, sich immer von Neuem quälen zu lassen. Dazu gibt er dem Menschen die Hoffnung: sie ist in Wahrheit das übelste der Übel, weil sie die Qual der Menschen verlängert.« (Nietzsche 1878).

Ist es nun ein Übel oder eine positive Entwicklung, wenn man erhofft, dass sich in der großen Menge an Daten mithilfe von Lesehilfen vielleicht neue Muster im großen Neumengetümmel zutage treten lassen? Wird schlussendlich nur gefunden werden, was gefunden werden will? Vielleicht. Und werden die großen Datenmengen den Blick für die Erkenntnis schärfen oder uns die Möglichkeit der Interpretation eher verstellen? Jedenfalls gehört es wohl zur Natur der Forscher, dass sie wie Pandora die Büchse öffnen wollen und werden, und seien darin auch noch so viele Neumen. – Es sei dieser augenzwinkernde Vergleich erlaubt, der insofern hinkt, da die Neumen natürlich keinerlei Übel darstellen, sondern einer der großen Schätze unseres Kulturerbes darstellen. Allein die große Menge an Notationszeichen rief den Vergleich hervor. Jedenfalls haben wir die begründete Hoffnung, dass aufgrund der Nutzung der Computertechnik der optischen Wiedererkennung in Bilddateien neue Forschungsmöglichkeiten entstehen. Denn ihre Anwendung wird den Forschern einmal ermöglichen, viele Handschriften und Notationen gleichzeitig zu untersuchen. Durch die Erweiterung der Datenmenge, die durchsucht werden kann, ist es möglich, neue Forschungsfragen zu kreieren und neue Methoden zu entwickeln für die Erforschung des Choralrepertoires. Die Bedeutung der einzelnen Notationszeichen wird anders als bisher erfasst werden können und es werden neue Erkenntnisse zutage

treten (z. B. hinsichtlich der Aufführungspraxis, des tonalen Systems, der Choralsemiologie und Exegese des Textes), weil zusätzlich zum heutigen Kenntnisstand messbare Daten gesammelt werden hinsichtlich des Vorkommens der Neume in der Handschriftenseite im jeweiligen Wort-Ton-Kontext. Zu einem späteren Zeitpunkt wird es möglich sein, dass Forscher verschiedenste Kombinationen von Neumen innerhalb des Repertoires suchen können und damit Muster erkennen, die Grundlage des mündlich tradierten Choralrepertoires sind. Forscher werden das Verhältnis von Melodieformel und zugehörigem Akzent einer Textsilbe besser beschreiben können, sowie besser verstehen, warum bestimmte Melodien oder Melodiebausteine und / oder Text oder Textbausteine im Repertoire an mehreren Stellen verwendet werden. Andere Forschungsfragen werden sich mit der Struktur des Gesangs beschäftigen, beispielsweise mit dessen Tonart, oder es werden dieselben Gesänge verglichen in Handschriften mit linienloser Notation und solchen, die Liniennotation enthalten. Überhaupt wird die Suche nach Melodien über die Grenzen von Handschriften hinweg unser Wissen über ihre Verbreitung des Repertoires in verschiedene Regionen Europas erweitern. Welche Handschriften dienten in welchen Klöstern als Vorlage für neue Codices? Die Suche in mehreren Handschriften wird ermöglichen, die Veränderung von Melodien über die Jahrhunderte hinweg zu studieren – so hinsichtlich des Wegfalls von Tönen in den Melodien oder des Auftretens von Melodievarianten. Aber auch über die Gestaltung der Rubriken sowie der Initialen und zur Buchillustration insgesamt werden weitere Erkenntnisse gemacht werden können. All dies wird zu einem vertieften Verständnis über die mittelalterlichen Skriptoria und die Gewohnheiten der Schreiber beitragen.

Bibliographie

Antiphonar Hartker. St. Gallen, Stiftsbibliothek, Cod. Sang. 390/391: Antiphonarium officii. <http://www.e-codices.unifr.ch/de/list/one/csg/0390>, <http://www.e-codices.unifr.ch/de/list/one/csg/0391>.

Corbin, Solange. *Die Neumen.* Köln: Volk, 1977.

DiMusEd: *Digitale Musik Edition / Digital Music Edition.* <http://www.dimused.uni-tuebingen.de/tuebingen_phase2.php>.

Engels, Stefan. »Adiastematische Neumen mit melodischer Zusatzbedeutung - ein wichtiges Hilfsmittel zur Melodierestitution.« *Beiträge zur Gregorianik (BzG)* 26 (1998). 63-80.

Engels, Stefan. »Neue Quellen zu Neumen mit adiastematischer Zusatzbedeutung in österreichischen Handschriften.« In Dobszay, László (Hrsg.). *Cantus Planus. Paper read at the 12th meeting of the Study Group, Lillafüred/Ungarn, 23.–28. August 2004.* Budapest: Inst. for Musicology of the Hungarian Acad. of Sciences, 2006. 455–470.

Floros, Constantin. *Universale Neumenkunde. Ursprung und Deutung der Lateinischen Neumen.* Kassel: Bärenreiter 1970.

Hornby, Emma und Rebecca Maloy. *Music and Meaning in Old Hispanic Lenten Chants. Psalmi, Threni and Easter Vigil Canticles.* Woodbridge: Boydell & Brewer, 2013.

MEI: *The Music Encoding Initiative.* <http://music-encoding.org>.

Nietzsche, Friedrich. *Menschliches, Allzumenschliches* I, *Ein Buch für freie Geister.* Chemnitz: Schmeitzner, 1878. Zitiert nach <http://www.textlog.de/21656.html>.

Old Hispanic Office: *Old Hispanic Office project.* <http://www.bristol.ac.uk/arts/research/oho-project>.

Pouderoijen, Kees und Ike de Loos. »Wer ist Hartker? Die Entstehung des Hartkerischen Antiphonars.« *Beiträge zur Gregorianik (BzG)* 47 (2009). 67-86.

SIMSSA: *Cantus Ultimus. The Single Interface for Music Score Searching and Analysis project.* <http://cantus.simssa.ca>.

Anhänge

Appendices

Kurzbiographien – Biographical Notes

Bernhard Assmann studierte an der Universität zu Köln Informationsverarbeitung, Mittlere und Neuere Geschichte und Historische Hilfswissenschaften. Danach betreute er das Digitalisierungsprojekt »Die Werke Friedrichs des Großen« an der Universitätsbibliothek Trier. Gegenwärtig ist er beim Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen in Köln beschäftigt.

Jennifer Bain, an Associate Professor of music at Dalhousie University, has published numerous articles on the music of Guillaume de Machaut and of Hildegard of Bingen, as well as on digital plainchant research and the reception of medieval music. She has edited a special early music issue of the Journal of Music Theory, and co-edited a multi-disciplinary collection of 18 essays on Guillaume de Machaut. Her book, Hildegard of Bingen and Musical Reception: the Modern Revival of a Medieval Composer, was published by Cambridge University Press in 2015.

Inga Behrendt is a research fellow at the Musicology Institute of the Eberhard Karls University Tübingen, and Professor »im Kirchendienst« of plainchant and hymnology at the Hochschule für Kirchenmusik der Diözese Rottenburg-Stuttgart. Her research concentrates on plainchant, especially neume notations, and on arrangements of early music for piano in publications by Hugo Riemann in the nineteenth and early twentieth century. She published articles on plainchant, the liturgical context of chant and medieval music and digital plainchant research.

Hartmut Beyer studierte Geschichte und Lateinische Philologie des Mittelalters und der Neuzeit in Göttingen und Münster. 2007 promovierte er mit einer Arbeit zu frühen neulateinischen Tragödien. Nach einer Zeit als Mitarbeiter im Münsteraner Exzellenzcluster »Religion und Politik« absolvierte er in Berlin und München das Referendariat für den höheren Dienst an wissenschaftlichen Bibliotheken. 2012 wurde er stellvertretender Leiter der Abteilung Forschungsplanung und Forschungsprojekte an der Herzog August Bibliothek, seit 2016 ist er stellvertretender Leiter der Abteilung Alte Drucke.

Bartosz Bogacz finished his bachelor's thesis in Computer Science at the University of Mannheim in 2010. Afterwards he joined the Heidelberg University and finished his master's thesis within the Database Research Group under the expertise of Prof. Michael Gertz. In 2014 he became a member of the Forensic Computational Geometry Laboratory (FCGL) at the Interdisciplinary Center for Scientific Computing (IWR), where he is working on his PhD project in the field of machine learning and word-spotting of cuneiform characters.

Hannah Busch studied German-Italian Studies in Bonn and Florence, and Textual Scholarship at the Free University of Berlin. Since 2013 she is member of academic staff at the Trier Center for Digital Humanities where she has been member of the eCodicology project from 2013 to 2016. Her research interests lie in the field of codicology, with a special focus on digital quantitative codicology and layout studies, and scholarly editing.

Alberto Campagnolo is the CLIR/DLF/Mellon Fellow for Data Curation in Medieval Studies at the Preservation Research and Testing Division of the Library of congress, Washington DC (2016-2018). He trained as a book conservator (in Spoleto, Italy - 2001) and has worked in that capacity in various institutions, including the London Metropolitan Archives, and the Vatican Library. He holds a PhD in Digital Humanities from University of the Arts, London (Ligatus Research Centre - 2015). He is especially interested in the digital representation of the physicality of books, and bookbindings in particular. He has served on the Digital Medievalist board since 2014, first as Deputy Director, and as Director since 2015.

Swati Chandna studied Computer Science (Master). She is a researcher at the Karlsruhe Institute of Technology (KIT). Her research interest are in the field of image processing, pattern recognition and information visualisation. From 2013 to 2016 she has been working in the eCodicology project.

Vincent Christlein received his diploma degree in Computer Science in 2012 from the Friedrich-Alexander-Universität (FAU) Erlangen, Germany. He is a doctoral student at the Pattern Recognition Lab, University of Erlangen-Nuremberg. His research interests lie in the field of computer vision, particularly in image forensics, and historical document analysis.

Erin Connelly is the CLIR-Mellon Fellow for Data Curation in Medieval Studies in the Schoenberg Institute for Manuscript Studies in the Kislak Center for Special Collections, Rare Books and Manuscripts. She holds a PhD in English from the University of Nottingham with a special interest in medieval medical texts and the relevance of medieval medicine for modern infections (Ancientbiotics). She is currently working on the first published edition of the fifteenth-century Middle English translation of Bernard of Gordon's Lilium medicinae.

Matthew Driscoll is senior lecturer in Old Norse Philology at the University of Copenhagen. His research interests include manuscript and textual studies, with special focus on popular manuscript culture in late pre-modern Iceland.

Franz Fischer is coordinator and researcher at the Cologne Center for eHumanities (CCeH), University of Cologne. Currently, he is coordinating the Marie

Skłodowska-Curie research and training programme DiXiT on digital scholarly editing. As founding member of the Institute for Documentology and Scholarly Editing (IDE) he is an editor of RIDE, a review journal on digital editions and resources, and editor-in-chief of Digital Medievalist.

Martin Gropp received a master's degree in Computer Science in 2012 from Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Germany. He worked as a researcher at the Computational Linguistics department of Saarland University between 2012 and 2015 and has recently joined the Pattern Recognition Lab of FAU as a doctoral student. His main research interest lies currently in bringing methods from language and speech processing to automatic handwriting analysis.

Svenja A. Gülden studierte Ägyptologie, Klassische Archäologie und Alte Geschichte an der Universität zu Köln. Seit 1996 war sie nacheinander wissenschaftliche Mitarbeiterin im DFG-Langzeitvorhaben »Edition des Ägyptischen Totenbuches« (Univ. Bonn), im Projekt »Multilingualism and Multiculturalism in Graeco-Roman Egypt« (Univ. Köln / Humboldt-Stiftung), und im DFG-Langzeitvorhaben »Die Nekropole von Assiut/Mittelägypten« (Univ. Mainz). Ihre Schwerpunkte liegen auf Handschriftenkunde und Funerärliteratur sowie dem Einsatz digitaler Medien in der Ägyptologie. Seit 2015 gehört sie zum Team des Mainzer Akademie-Vorhabens »Altägyptische Kursivschriften«.

Philipp Hegel graduated in Comparative Literature, History, and Philosophy at Bielefeld University, and in Textual Scholarship at Free University of Berlin. Since 2011 he is a Research Assistant at Darmstadt University of Technology. Currently he is a member of the Collaborative Research Center 980 »Episteme in Bewegung«.

Kate Helsen currently teaches in the Music Research and Composition department at the Faculty of Music at Western University. Before this, she held a post-doctoral fellowship at the University of Toronto. Her doctoral research, under David Hiley at the University of Regensburg, focused on the Great Responsory repertory in the Gregorian tradition. She is published in Plainsong and Medieval Music, Acta Musicologica, the Journal of the Alamire Foundation, SPECTRUM, and Early Music. She has been a researcher with musicological projects in Germany, Portugal, Ireland, and Canada.

Gábor Hosszú is associate professor at the Budapest University of Technology and Economics, where he has been awarded by the title of doctor habil and venia legendi in the field of Informatics, Engineering and Technology in 2013. He received M.E. degree in electrical engineering at the Technical University of Budapest in 1985, the Academic degree of Technical Sciences (Ph.D.) in 1992, and

MSc in Law at the Pázmány Péter Catholic University in 2011. His interests are mainly in the areas of Internet-based communication, statistical evaluation of bioelectronic signals, and computational paleography.

Celia Krause studierte Altertumswissenschaften in Heidelberg und Köln. Seit 2011 ist sie wissenschaftliche Mitarbeiterin in verschiedenen Projekten am Institut für Sprach- und Literaturwissenschaft der TU Darmstadt. Ihre Interessensgebiete liegen in den Bereichen digitale Grundlagenforschung, Datenmodellierung, Bild-Text-Relationen und Neue Medien. Sie ist im Akademienprojekt »Altägyptische Kursivschriften« tätig.

Andreas Maier graduated in 2005 in Computer Science and received his PhD in 2009. He was one of the pioneers in medical speech processing. Since 2009 he started working on image processing with stays at Stanford University and Siemens Healthcare. Since 2012 he returned to the University of Erlangen-Nuremberg and heads the Pattern Recognition Lab since 2015.

Hubert Mara studied Computer Science at Vienna University of Technology in Austria. He was a Marie-Curie fellow in the Cultural Heritage Informatics Research Oriented Network (CHIRON, FP7) at the University of Florence in Italy in 2007 and 2008. In 2009, he became a member at the Interdisciplinary Center for Scientific Computing (IWR) of the Heidelberg University, Germany, where he finished his PhD in 2012 within the Visualization and Numeric Geometry Group (VNGG). Since 2014 he is an independent research group leader and founder of the Forensic Computational Geometry Laboratory (FCGL) at the IWR.

Jörn Münkner hat sich 2008 an der HU Berlin promoviert. Bis 2011 koordinierte er das bilaterale Promotionsnetzwerk »PhD-Net: Das Wissen der Literatur« an der HU. 2011 wechselte er an das Institut für Germanistik an die Universität Kassel. Seit November 2015 ist er in dem an der Herzog August Bibliothek Wolfenbüttel durchgeführten Projekt »Frühneuzeitliche Gelehrtenbibliotheken« angestellt; das Projekt ist Teil des Forschungszusammenhangs »Autorenbibliotheken« im Forschungsverbund Marbach Weimar Wolfenbüttel (MWW).

Dot Porter is Curator, Research Services in the Kislak Center for Special Collections, Rare Books and Manuscripts at the University of Pennsylvania. She holds an MA in Medieval Studies from Western Michigan University and an MS in Library Science from UNC-Chapel Hill. She is the lead researcher on VisColl.

Dariya Rafiyenko is a research fellow at the project »eXChange«, University of Leipzig, and PhD student at the University of Cologne, Department of Byzantine Studies. She has completed a digital edition of the anonymous fragments of an

early Byzantine historiographer (the so-called Anonymus post Dionem, commonly dentified with Peter the Patrikios). Her further research interests include corpus linguistics and extraction of semantic data from the corpus of Ancient Greek and Byzantine texts.

Andrea Rapp hat Germanistik, Kunstgeschichte und Ethnologie an der Universität Trier studiert. Nach Stationen an der Universität Trier und der Niedersächsischen Staats- und Universitätsbibliothek ist sie seit 2010 Professorin für Germanistik – Computerphilologie und Mediävistik an der Technischen Universität Darmstadt, seit 2017 Vizepräsidentin für wissenschaftliche Infrastruktur ebendort. Die Forschungsinteressen und Arbeitsgebiete liegen in den Bereichen der Buch- und Bibliotheksgeschichte, der historischen Schreibsprachgeschichte, von illustrierten mittelalterlichen Handschriften, digitaler Editionsphilologie, Lexikographie, Virtuellen Forschungsinfrastrukturen sowie allgemein in den Digital Humanities.

Nanette Rißler-Pipka is Private Lecturer at the University of Siegen. In 2016/17 Acting Professor for Romance Literature at the Catholic University of Eichstätt-Ingolstadt. Member and co-founder of the research network on Spanish and Hispanic cultural magazines at the University of Augsburg (www.revistas-culturales.de). Her research interests focus on text-image-analysis in literary and cultural contexts, intermediality and digital humanities. Postdoctoral thesis (Habilitation) on Picasso's writings: *Picassos schriftstellerisches Werk: Passagen zwischen Bild und Text* (Bielefeld 2015).

Patrick Sahle is »außerplanmäßiger« professor for Digital Humanities at the University of Cologne. He works for the Cologne Center for eHumanities (CCeH) and cares for several research projects in the wide field of Digital Humanities. His interest in digital codicology and palaeography goes back to his original education in medieval history and auxiliary sciences (UoC, Vatican Library) and various research projects on manuscripts and other kinds of documents from archives and libraries. As a founding member of the IDE he has been involved in the making of Codicology and Palaeography in the Digital Age vols. 1-3.

Torsten Schaßan hat Geschichte, Germanistik und Philosophie an der Universität zu Köln studiert. Er ist wissenschaftlicher Mitarbeiter an der Abteilung Handschriften und Sondersammlungen der Herzog August Bibliothek Wolfenbüttel. Er betreut dort die digitale Erschließung der historischen Bestände, insbesondere die Handschriftenkatalogisierung und digitale Editionen. Seine Arbeits- und Forschungsschwerpunkte sind die Digitalisierung des kulturellen Erbes, Datenstrukturen und Metadatenformate sowie Methoden der Digital Humanities.

Katrin Schmidt hat an der Hochschule Hannover Informationsmanagement mit dem Schwerpunkt auf wissenschaftliche Bibliotheken studiert. Seit 2010 ist sie an der Herzog August Bibliothek Wolfenbüttel (HAB) tätig und hat dort die Digitalisierungsprojekte Helmstedter Drucke Online sowie VD17 Mainstream bibliothekarisch betreut. Seit 2014 ist sie Bibliothekarin im Forschungsverbund Marbach Weimar Wolfenbüttel (MWW), im Projekt »Autorenbibliotheken«. Seit 2016 ist sie zusätzlich mit bibliothekarischen Aufgaben im Bereich der Wolfenbütteler Digitalen Bibliothek (WDB) betraut.

Timo Steyer ist wissenschaftlicher Mitarbeiter im Bereich Digital Humanities im Forschungsverbund Marbach Weimar Wolfenbüttel (MWW). Zuvor war er in verschiedenen Projekten an der Herzog August Bibliothek Wolfenbüttel beschäftigt, u.a. In »AEDit-Frühe Neuzeit«. Seine Forschungsschwerpunkte liegen in den Bereichen digitales Publizieren, digitale Editionen und Datenmodellierung.

Ursula Verhoeven wurde an der Universität zu Köln im Fach Ägyptologie promoviert und habilitiert und ist seit 1998 Universitätsprofessorin an der Johannes Gutenberg-Universität Mainz. Sie ist Leiterin (seit 2010 Ko-Leiterin) des DFG-Langzeitvorhabens »Die Nekropole von Assiut/Mittelägypten« und seit 2015 Leiterin des Mainzer Akademie-Vorhabens »Altägyptische Kursivschriften«. Ihre Schwerpunkte sind die altägyptische Schriftkultur, insbesondere hieratische Texte des Neuen Reiches und der Spätzeit, außerdem altägyptische Literatur, Religion und Funerärkultur.

Enrique Vidal is a full professor of computer science in the Universitat Politècnica de València (Spain) and former co-leader of the PRHLT research center in this University. He has published more than two hundred research papers in the fields of Pattern Recognition, Multimodal Interaction and applications to Language, Speech and Image Processing and has leaded many important projects in these fields. Dr. Vidal is a member of the IEEE and a fellow of the International Association for Pattern Recognition (IAPR).

KPDZ 1 – CPDA 1

Kodikologie und Paläographie im Digitalen Zeitalter / Codicology and Palaeography in the Digital Age, hg. v. Malte Rehbein, Patrick Sahle und Torsten Schaßan unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009.
ISBN 978-3-8370-9842-6
Online: <http://kups.ub.uni-koeln.de/volltexte/2009/2939/>
Der gedruckte Band kann zum Preis von € 49,- über den Buchhandel, über amazon.de und über die Webseite des Verlages bezogen werden:
<http://www.bod.de/index.php?id=1132&objk_id=217805>.
You can order the printed version at the price of € 49,- from your local bookstore, via amazon.de or via the website of the publishing house:
<http://www.bod.de/index.php?id=1132&objk_id=217805>.

Beiträge – Contributions

KPDZ 2 – CPDA 2

Kodikologie und Paläographie im Digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2, hg. v. Franz Fischer, Christiane Fritze und Georg Vogeler unter Mitarbeit von Bernhard Assmann, Malte Rehbein und Patrick Sahle. Schriften des Instituts für Dokumentologie und Editorik 3. Norderstedt: Books on Demand, 2010.
ISBN 978-3-8423-5032-8
Online: <http://kups.ub.uni-koeln.de/4337/>
Der gedruckte Band kann zum Preis von € 59,- über den Buchhandel, über amazon.de und über die Webseite des Verlages bezogen werden:
<http://www.bod.de/index.php?id=296&objk_id=477356>.
You can order the printed version at the price of € 59,- from your local bookstore, via amazon.de or via the website of the publishing house:
<http://www.bod.de/index.php?id=296&objk_id=477356>.

Beiträge – Contributions

KPDZ 3 – CPDA 3

Kodikologie und Paläographie im Digitalen Zeitalter 3 / Codicology and Palaeography in the Digital Age 3, hg. v. Oliver Duntze, Torsten Schaßan, und Georg Vogeler, unter Mitarbeit von Bernhard Assmann, Johanna Puhl und Patrick Sahle. Schriften des Instituts für Dokumentologie und Editorik 10. Norderstedt: Books on Demand, 2015.
ISBN 978-3-7347-9899-3
Online: <http://kups.ub.uni-koeln.de/7151/>
Der gedruckte Band kann zum Preis von € 29,- über den Buchhandel, über amazon.de und über die Webseite des Verlages bezogen werden:
<http://www.bod.de/buch/oliver-duntze/kodikologie-und-palaeographie-im-digitalen-zeitalter-3/9783734798993.html>.
You can order the printed version at the price of € 29,- from your local bookstore, via amazon.de or via the website of the publishing house:
<http://www.bod.de/buch/oliver-duntze/kodikologie-und-palaeographie-im-digitalen-zeitalter-3/9783734798993.html>.

Beiträge – Contributions