Versioning Cultural Objects
Digital Approaches

Schriften des
Instituts für Dokumentologie und Editorik

herausgegeben von:

Band 13

Schriften des Instituts für Dokumentologie und Editorik — Band 13

Versioning Cultural Objects
Digital Approaches

---

edited by

Roman Bleier, Sean M. Winslow

2019

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 18. Dezember 2019.

2019

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7504-2702-0
Einbandgestaltung: Julia Sorouri and Stefan Dumont; Coverbild gestaltet von Vinayak Das Gupta.
Satz: Roman Bleier, Sean M. Winslow und LuaTEX

Contents

Electronic Versioning

Appendix

Preface

The present volume, *Versioning Cultural Objects: Digital Approaches*, is a collection of selected essays that were first presented and discussed at a symposium at An Foras Feasa, The Research Institute for the Humanities at Maynooth University, in December 2016. The idea of the volume is to start a discussion about the different types of versions we are dealing with in the digital humanities (texts, objects, analogue, and digital resources) across disciplines.

The editors of the volume are grateful to the Digital Arts and Humanities structured PhD programme funded by the Irish government's Programme for Research in Third-Level Institutions (PRTLI) Cycle 5 for funding the author's symposium as a space where interesting, cross-disciplinary discussion happened. Special thanks are due to Susan Schreibman and Vinayak Das Gupta and the staff of An Foras Feasa for organising and hosting the symposium. Vinayak Das Gupta was originally co-editor of the volume and his contributions and support were crucial to the early stages of this publication, including the selection of authors and peer-reviewers. We are very grateful for his support during the development of this work and for designing the cover image. Many thanks go also to Bernhard Assmann (Cologne) and Patrick Sahle (Wuppertal) for helpful suggestions and advice during the typesetting process, to Julia Sorouri (Cologne) and Stefan Dumont (Berlin) for the design of the cover. Last but not least, thanks go to the Institute for Documentology and Scholarly Editing (IDE) for its continued support during the editing process and to the peer-reviewers for their helpful comments and critical advice.

Graz, December 2019, the editors

Introduction

Introduction: Versions of Cultural Objects

Roman Bleier, Sean M. Winslow

Abstract

The *version* of a cultural object is identified, defined, articulated, and analysed through diverse mechanisms in different fields of research. The study of versions allows for the investigation of the creative processes behind the conception of the object, a closer inspection of the socio-political contexts that affect it, and may even provide the means to investigate the object's provenance and circulation. At a symposium at Maynooth University, scholars from different research areas exchanged ideas about different forms of media, including text, image, and sound, to work towards an understanding of the term *versioning* in the broadest sense. While the understanding of versions and related terminology differs between disciplines, a cross-disciplinary dialogue will highlight the range and depth of existing studies and provide an interdisciplinary understanding of the term versioning which will be useful for a more holistic conceptualisation. The present volume tries to contribute to this dialogue by providing eight peer-reviewed articles resulting from the discussion and presentations held at Maynooth University.

The breadth and applicability of the concept of a *version* is at the core of this volume. Questions like: *Can the word* version *be applied uniformly across disciplines? Does the meaning of the word change?* drove the editorial decisions in bringing together the various participants in the original symposium in Maynooth which was the beginning of this volume. The range of the answers presented here underline the striking multivariance of the term, and the way that different humanities researchers are using it, from music to genetic criticism to versioning as it is understood in the management of shared code databases. By choosing these articles, we hope that we can offer not only a sense of the range of the field, but invite the reader to think about the many facets that have to be considered in order to fully understand the semantic lifting done whenever the word version is encountered, and how we might begin to form a shared understanding of the fullness of the term, but also where it needs more support and specificity.

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, V–XI.

1 The genesis of this volume

This volume had its genesis in the work of An Foras Feasa during 2015 and early 2016, The Research Institute for the Humanities at Maynooth University, then headed by Susan Schreibman. Roman worked on the redesign and release of the *Versioning Machine 5.0*, a publication framework for the display and visual analysis of multiple versions of a text. Vinayak worked on a theoretical framework to capture electronic metadata of visual resources (see Das Gupta); it was in that setting that the question arose of how to record reproductions and the context they were produced in. In order to foster an interdisciplinary discussion about the topic, they organized a symposium as a platform to present and discuss the various disciplinary approaches. In addition to the presentation of papers, the participants worked in groups to examine related terminology. This cross-disciplinary exchange can be seen in the finished chapters.

2 Why was the term *versioning* used?

The term versioning is more frequently used in the context of software versioning and electronic version control. The *Versioning Machine*, developed by Schreibman et al. in the early 2000s (launched in 2002), introduced the term in the sense of *exploring variation between textual versions of a work* into the digital humanities community (see Schreibman, "Re-Envisioning"; Schreibman et al., "The Versioning Machine"). With the *Versioning Machine*, Susan Schreibman investigated the composition process of Thomas MacGreevy's poetry by comparison and parallel reading of various versions of the poems.

Taking Schreibman's work as a point of departure, and the attendant realization that versioning means different things to different disciplines and to different practitioners, the edited articles in this volume illustrate the range and depth of existing studies of versions and will (we hope) provide a first step towards a platform for an interdisciplinary discussion and understanding of the concept. The volume engages with versioning in the digital humanities in three primary areas: the conceptualisation of versions in different humanities disciplines, the methods involved in the electronic modelling of versions of cultural objects, and the representations of digital versions. Individual articles may cover one or more of these areas in varying depth. Appropriately enough for a book on versioning, our volume opens with Elisa Nury's dissection of the meaning of *variant reading* in textual scholarship. She asks whether the concept of "variance" means the same thing in different disciplines, emphasizing the importance of contextualisation of the term and presents an implementation of a digital representation of a *reading*, which is a first step to conceptualise variant reading, using the CollateX JSON data format.

3 Textual versions and digital editing

The advent of the digital age has led to a profusion of digital versions of documents, but problems in dealing with versions are hardly new: palaeographers had to deal with different versions of scripts, numismatists with versions of coins, archaeologists for instance with marble versions of Greek bronze statues or motives on Greek red-figure pottery, textual scholars with versions of written sources, art historians with different versions of artworks. The methodologies developed in a pre-digital context still have validity today and many scholarly discussions have continued and are being adapted in the digital scholarly context. So what does it mean when we, as digital humanists, talk about versions? Where do traditional approaches of the pre-digital age end, and what do new, digital approaches entail?

For instance, digital textual editing discussions about versions go in different directions: in stemmatology and copy-text editing, an editor has to establish which variation between different manuscript witnesses to "trust" in order to establish a "safe text" that comes as close as possible to an author's original work. Editors following the genetic editing approach try to untangle the various layers of revisions and changes made to a manuscript over time. While genetic editing was the exception in print, the flexibility of the digital medium to represent different layers of a text has led to a substantial increase in the development of such editions (see Pierazzo, "Digital Documentary Editions"). One of the central characteristics of digital scholarly editing is the separation of data and presentation. The data is usually represented using the standard of the Text Encoding Initiative (TEI) which allows the modelling of versions of a text in concordance with the traditional editing approaches (see Burghart; Pierazzo, "Facsimile").

Three chapters in the volume, by Martina Scholger, Richard Breen, and Christian Thomas present case studies of digital scholarly editing projects that investigate different kinds of versions and variance.

Martina Scholger's chapter, "Pieces of a Bigger Puzzle," explores her work on a digital scholarly edition of the notebooks of Hartmut Skerbisch, an Austrian visual artist. The notebooks contain a network of references to music, literature, and other visual art works as well as numerous sketches, constructional drawings, and diagrams of Skerbisch's installations in various stages of conceptual planning. Lacking the finished installations, Scholger uses a genetic criticism approach to uncover and identify the various versions which are the result of the artist's creative process and to examine the relationship between sketch and visitors' reports of the final installation in relation to the genesis of his artistic work.

Richard Breen explores the transmission of the many variants of "The Unfortunate Rake." We might wonder what "St. James Infirmary Blues" has in common with "Streets of Laredo," a nineteenth-century cowboy song, or what either has to do

with an Irish folk song. To explore and visually show the motivic similarities, Breen uses the *Versioning Machine* and StoryMapJS to map and narrate the distribution of the song variations across the globe. Versions in this case are similar motives that developed across variations of the song, uniting a seemingly-disparate corpus in one family network.

Like Scholger, Christian Thomas seeks to reconstruct a missing event, in this case, Alexander von Humbolt's Kosmos-lectures from Humbolt's fragmentary manuscripts, lecture notes taken by attendees, and related documents. These fragmentary reports about the lectures can be viewed as witnesses or versions that can enrich and complement our knowledge of the lectures and its contents, but they may also present conflicting information and narratives. The question is how to deal with such a rich and diverse number of primary sources, especially if—like in the case of the lecture notes—their authorship and origin is not always clear.

4 Considering other representational forms as versions

Historians have for many decades made editions where *regest*s, short abstracts listing the main information about a text, have been used for extracting and summarising information important for historical research. Focusing on content—rather than wording—allows the creation of versions of texts enriched by external information (in the form of RDF) in order to to find connections and support advanced search functionalities goverened by a conceptual model (see Vogeler). Consequently, as representations of the information layer, abstracts, regests, or metadata should be considered as expressing a version of the same, each supplementing or replacing other possible versions based upon project and disciplinary needs. In this volume, Georg Vogeler uses the example of medieval charters to discuss copies and what other kind of versions were added in the digital world: transcriptions and reproductions of a charter in print and digital form, archival and scholarly descriptions, and metadata become part of his model. He suggests a graph-based data model with RDF that allows a more flexible and suitable approach than current XML or relational database solutions.

This focus on the information layer also leads one to think about the various trajectories in the production and commodification of an object, and the meanings and values associated throughout these histories, which can be referred to as *object biographies* (see Kopytoff). Treating the history of objects as a version of what they are, fully in parallel with their content, reminds us that objects and texts, as they come down to us today, may not only exist in different versions (as of, say, a painting), but are usually different versions of themselves, having undergone changes, whether

physical or in terms of their recontextualization, which affect our interpretation, and are themselves, in effect, variant readings of the object. That these changes should have a temporal aspect which needs to be considered is no surprise, and Athanasios Velios and Nicholas Pickwoad's contribution to this volume, presenting CIDOC-CRM events for reconstructing the history of binding structures, attempts to address the need to formally document this temporal aspect in digital codicology.

5 Electronic texts and version control

The book concludes with two papers discussing principles of the versioning of electronic documents, comparing versions of electronic documents, and problems when trying to collaboratively work with documents in an online environment. Specifically-electronic considerations for editing should be taken into account, as in Gioele Barabucci's exploration of different abstraction levels of electronic documents which can be described by their content, model, variants, and physical embodiment. The paper describes the problem and presents a formal solution in the CMV+P model. The implementation of this model would enable a user (human or computer) to precisely describe and communicate the type of version of an electronic document the user is interested in. Practical applications include document comparison tools which could operate on a CMV+P based model to compare only the levels of primary interest.

Metacontextual issues around project management and collaboration are considered by Martina Bürgermeister, who discusses the importance of versioning control systems for digital collaboratory research environments. She critically analyses collaborative projects such as Annotated Books Online, Monasterium.net and Wikipedia by exploring how collaboration and versioning control is implemented by these organizations. She concludes that existing collaboratories do not satisfy the needs of humanities research, and suggests conceptual models which will help us to classify the various types of changes happening in electronic documents during collaborative work and the relationships between them.

6 Concluding remarks

In a way, a single thread connects Nury's opening article, which starts the volume with a solid grounding in the text critical concept of versions, to Bürgermeister's closing article, which deals with versioning metadata from current development practices. We proceed from the big, basic question of what a version is, through case studies, to domain-specific formal systems for representing knowledge (discussed by Vogeler, Velios and Pickwoad), down to a narrowed and focused exploration of the actual codepoints that represent word information (Barabucci). Another thread

might go from Nury's analogue context, through case studies by Breen, Thomas, and Scholger—which all present material which could be represented in an analogue edition, but where digital methods help to present the complexity of the data more clearly than was possible in print—to the purely-digital representations enabled by graph data and the digital structure of word data itself, and the data about data that is collected by a versioning system. Yet another thread would wind in a convoluted and hopelessly knotted fashion, detouring for all the similarities among the articles. As an example, in both Scholger's work on artists notes and Thomas' work on the Kosmos-lectures, we lack a direct and authoritative version of the "main event" (the installation for Skerbisch and the lectures for von Humboldt), leading to a reconstructed *ur*-version which is itself unstable and subject to variance in interpretation. Here, we see techniques developed for critical textual editing applied to the reconstruction of performance. Vogeler's work on charters highlights similar issues for drafts, as the final, legal version of a charter can be preceded by non-legal drafts, and followed by promulgations and re-issues that are separate legal acts of the same basal charter text. Here, the versions speak to both the textual development of the charter as well as the various instances of its legal effectuation. We hope that these examples will encourage people to give thought to how the concept of versioning changes and with what kind of new versions we are dealing in a digital context.

Bibliography

Burghart, Marjorie. "Textual variants." *Digital Editing of Medieval Texts: A Textbook*, edited by Marjorie Burghart. 2017. www.digitalmanuscripts.eu/digital-editing-of-medieval-texts-a-textbook/. Accessed 21 Aug. 2019.

Das Gupta, Vinayak. "Albums in the attic." *An investigation of photographic metadata.* Studia Universitatis Babes-Bolyai 62. 2017, pp. 57-74.

Kopytoff, Igor. "The cultural biography of things: commoditization as process." *The Social Life of Things: Commodities in Cultural Perspective*, edited by Arjun Appadurai, Cambridge University Press, 1986, pp. 64-91.

Schreibman, Susan, et al. "The Versioning Machine." *Literary Linguistic Computing*, vol. 18, no. 1, 2003, pp. 101–7.

———, editor. *The Thomas MacGreevy Archive.* 2007. www.macgreevy.org/. Accessed 21 Aug. 2019.

———. "Re-Envisioning Versioning A Scholar's Toolkit." *Digital Philology and Mediaeval Text,* edited by Ciula Arianna and Francesco Stella, 2007, pp. 93-102.

———, et al. *Versioning Machine 5.0.* 2016. v-machine.org/. Accessed 21 Aug. 2019.

Vogeler, Georg. "The 'assertive edition'." *International Journal of Digital Humanities*, vol. 1, 2019, pp. 309–22. doi.org/10.1007/s42803-019-00025-5. Accessed 12 Oct. 2019.

Pierazzo, Elena. "Digital Documentary Editions and the Others." *The Annual of the Association for Documentary Editing*, vol. 35, 2014. scholarlyediting.org/2014/essays/es-

say.pierazzo.html. Accessed 21 Aug. 2019.
———. "Facsimile and Document-Centric Editing." *Digital Editing of Medieval Texts: A Textbook*, edited by Marjorie Burghart. 2017. www.digitalmanuscripts.eu/digital-editing-of-medieval-texts-a-textbook/. Accessed 21 Aug. 2019.

What is Variance?

Towards a Model of (Variant) Readings

Elisa Nury

Abstract

In scholarly editing, more particularly in the context of collating various versions of a text, the definition of a variant reading is crucial. Yet, despite its importance, the meaning of a variant reading is often reduced to a "difference." The reason for such a vague definition is that what makes a variant can largely depend on the field of study: scholars of the Homeric oral tradition will consider different variants from scholars of medieval traditions or early printed texts, or from genetic critics. This contribution will focus on the modelling of a *reading*, arguing that formalizing this concept is necessary in order to define, and thus model, a *variant*. This article will also address digital representation of a reading by focusing on one implementation: the JSON data format used in conjunction with collation programs such as CollateX.

What is a version? In textual criticism, the term *version* may specifically describe a major rewriting of a work, possibly by the author. Here, however, we will consider versions in a broader sense. The critical comparison—or collation—of different versions of one text is a necessary step during the preparation of a text-critical scholarly edition. Each version of the text is recorded in a document—or witness—and consists of readings, i.e., the particular word or words found at a given point in the text. In this context, a version is determined, amongst other characteristics, by the *differences* in the words found in the text, or *variant readings*. Variant readings are important since they provide valuable information regarding how versions are related to each other and how the text evolved through transmission. This article will focus on the modelling of *readings*, arguing that formalizing this concept is necessary to define, and model, variant readings. We will show how reading was a technical term that was used quite consistently through the ages, until it was defined with precision. Then we will establish the basis for a model by selecting important features of textual readings according to the previously examined definitions. These features, such as the textual content (or absence thereof), its size, and location in the text, will be discussed, raising various issues. This article will also address digital representation of a reading by focusing on one implementation: the JSON data format used in conjunction with collation programs such as CollateX. As we will see, the concept of variant readings may depend on the tradition of the text in consideration, and a variant in Homeric epic is different from a variant in a medieval tradition. The concept of variant is also dependent on the purpose of the comparison: a scholar attempting to reconstruct a

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 3–23.

stemma, or a linguist, may need to examine different variants. Therefore, a model of a reading should make it possible to distinguish different sets of variants depending on the context, and we will examine how the JSON implementation makes it possible with a few examples.

Let us consider the example of figure 1, where four versions of a sentence are aligned. When comparing the sentences of A, B C,and D, some readings can be considered equivalent in all four sentences, such as *The* or *upon*; other readings are different and change the meaning of the sentence: the absence of the adjective *bright* in sentence B, the triplet *star/sun/stars*, and the verbs with different tense (*shines* and *shone*). Finally, some readings are different, but may not alter the sense of the sentence (such as *worlde* and *world* or *sun* and *sunne*). Readings are thus divided between equivalent readings and different readings, and among the different readings a set of readings may be considered significant variant readings (see figure 2).

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 1: Readings.

In the short collation extract of figure 1, there are four places where differences appear in the text. However, not all differences between readings are necessarily considered variant readings in any possible context. Scholarly opinions on this point range widely: from the view that every difference is a variant (Andrews) to considering only a limited number of "significant" differences to be variants, for instance, in the context of New Testament criticism, and therefore it is not enough to define a variant simply as a difference:

> The common or surface assumption is that any textual reading that differs in any way from another reading in the same unit of text is a "textual variant", but this simplistic definition will not suffice. Actually, in NT textual criticism the term "textual variant" really means—and must mean—"*significant*" or "*meaningful* textual variant" (Epp 48).

In fact, the concept of *variance* has evolved with time and according to several theories. Since the nineteenth century, many scholars contributed to the development of a

method for the establishment of genealogical relationships between manuscripts: the so-called Lachmann method. Maas in particular focused on a specific category of differences: shared errors, or indicative errors, can be used as a guide in order to assess the witnesses of the text and determine their relationships into a *stemma codicum*, or genealogical tree of textual witnesses.[1]

Greg separated variant readings into accidental and substantial, following the idea that some differences (substantials) have more importance than others (accidentals):

> [W]e need to draw a distinction between the significant, or as I shall call them, *substantive* readings of the text, those namely that affect the author's meaning or the essence of his expression, and others, such in general as spelling, punctuation, word-division, and the like, affecting mainly its formal presentation, which may be regarded as the accidents, or as I shall call them, *accidentals* of the text (Greg 21).

In the twenty-first century, scholars started to compare textual variants to DNA mutations and applied concepts from evolutionary biology and phylogenetics to textual criticism (Barbrook et al.; Salemans; Heikkilä). Lastly, in opposition to the distinction between accidental and substantial variants, Andrews suggested a big data approach where every difference is a variant.

With the introduction of Lachmann's method, shared errors became the object of scholarly attention, and much work was done on the description and classification of the kind of errors committed by scribes who were copying manuscripts by hand. The cause of the error, as well as its conscious or unconscious character, is generally taken into account. Since the conscious modifications of scribal corrections were often attempts at improving or restoring the text, the terms *innovation* and *secondary reading* are frequently preferred to *error*. One of the most comprehensive review of errors was published by Havet, but other scholars have proposed other typologies of errors (Petti; Love; Reynolds and Wilson). These typologies often divide errors into four types: additions, omissions, substitutions and transpositions (Petti). When the scribe is consciously modifying the text, Petti (28–29) refers to scribal corrections as insertions, deletions and alterations instead of additions, omissions and substitutions. In parallel, many fields of study have offered their own definitions for variants according to their needs and their perspective on the text. From oral traditions such as Homeric epic to early printing, from medieval traditions to genetic criticism, from linguistics to phylogenetics, variants take many forms depending on the context: *multiformity* (Nagy), *early* or *late* states (Dane), variants at the sentence level (Cerquiglini), *open* variants, *type-2* variants (Salemans), and so on. The task of proposing a model for

[1] Witnesses are documents which bear a copy of a text, and may be either manuscripts or printed editions. The stemma is a diagram that represents the relationships between those witnesses.

variant readings which would be suitable in any of the possible contexts, seems at best challenging, if not impossible. Rather than dealing directly with variants, this article will focus on modelling readings, especially textual readings. Not all readings are variant readings, but variants are always readings which differ in some respect from one another (see figure 2). Once readings have been modelled, variant readings could be more easily modelled as a set of readings, with various criteria according to each discipline (V1, V2, V3). However, modelling those subsets will not be in the scope of this article. In order to propose a model for readings, we will first review the origins and usage of the term as well as its definitions in Section 1. The analysis of definitions will provide a first outline for a model, which will be discussed in Section 2.

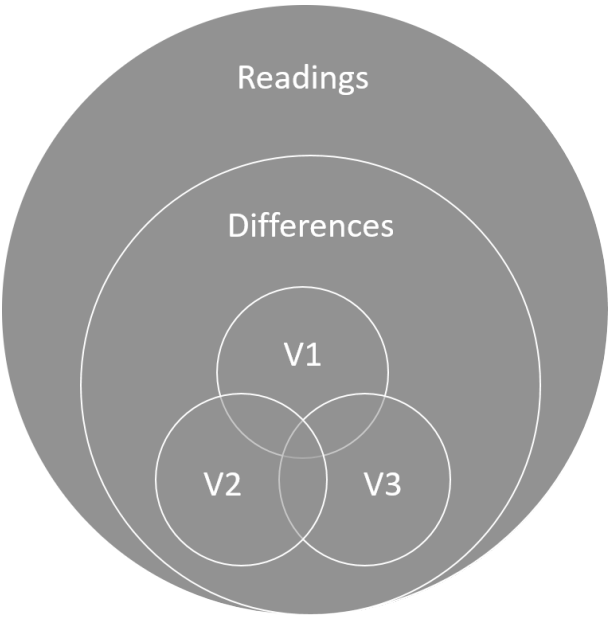


Figure 2: Readings, differences and variants.

1 Readings in context

Reading is a technical term that has long been used in the context of textual criticism and philology. It was already attested with Alexandrian critics: terminology included *graphe* (what is written), and *anagnosis* (what is read, a reading). The Latin equivalents are *scriptura* and the most common *lectio* (Montanari 26). The terms used by scholars

of Antiquity imply a distinction between the words that are actually written on the page as opposed to the interpretation of the text. In English as well, a reading implies a form of interpretation; it could be read in more than one way. Here are a couple of examples where the words *scriptura* and *lectio* are used to qualify textual variation:

> *Obolus, id est, virgula iacens, adponitur in verbis vel sententiis superflue iteratis, sive in his locis, ubi* ***lectio*** *aliqua* ***falsitate*** *notata est, ut quasi sagitta iugulet supervacua atque falsa confodiat.* Isidore 1.21.3.

> The obelus, that is, a horizontal stroke, is placed next to words or sentences repeated unnecessarily, or by places where some **passage** is marked as **false**, so that like an arrow it slays the superfluous and pierces the false. (Barney et al.)

> *"Et idcirco inportunissime," inquit, "fecerunt, qui in plerisque Sallusti exemplaribus* ***scripturam*** *istam* ***sincerissimam*** *corruperunt."* Aulus Gellius 20.6.14.

> "And therefore," said he, "those have acted most arbitrarily who in many copies of Sallust have corrupted a **thoroughly sound reading**." (Rolfe)

Here the nouns *scriptura* and *lectio* have been emphasized, as well as the term which qualifies them. As these passages demonstrate, there was a strong focus in Antiquity on whether a reading is corrupt or sound. When producing a new literary book, Hellenistic scholars used to correct a single copy of a work, instead of comparing as many copies as possible as modern editors do. This practice led Hellenistic scholars to become correctors of a specific work, and some experts compared them to editors (Montanari). Therefore, the need to distinguish between authentic and spurious readings arose, which may have motivated the dichotomy between sound versus corrupt readings, true versus false. The concept of variant reading, however, appeared much later during the Renaissance. In the Renaissance, Humanist scholars who were rediscovering and editing classical texts of Latin and Greek literature started to deploy technical terms that would become the base of the language of textual criticism. Silvia Rizzo's *Lessico Filologico degli Umanisti* provides invaluable information about the vocabulary in use amongst famous Humanists in the fourteenth and fifteenth centuries. By analysing their correspondence and publications, Rizzo was able to extract global definitions and explain what they meant when they used a given word. During the Renaissance, as Rizzo (209–13) shows, *lectio* and *scriptura* continued to be used as synonyms in much the same way as in Antiquity, for a passage of a text that can be read in a manuscript or an edition. Renaissance scholars would apply the term to readings from manuscripts as well as conjectures by other Humanists, and would

mostly describe those readings as either correct (*recta, sincera*) or incorrect (*corrupta, mendosa*) according to their judgement. At the same time, the concept of *variant reading* started to be used more precisely with *varietas* (diversity) and in expressions where *lectio* or *scriptura* were used in connection with the adjective *varius*. Lorenzo Valla and Girolamo Avanzi have both used *varia lectio* and *varia scriptura* to describe a portion of text with different possible readings, as reported by Rizzo (213). Valla was accused by Poggio of having presumptuously corrected a verse from Sallustius' first Elegy. Valla replied to Poggio that he did not emend Sallustius but merely chose one reading in a passage that varies (*varia scriptura*), even though the reading was attested only in very few manuscripts.[2] Another scholar, Avanzi, was asked for his opinion on a difficult passage from Catullus I, 9. He offers no solution of his own to emend the corrupted text, but he sends to his correspondent a list of conjectures (*varia lectio*) proposed by others.[3]

The usage of *lectio* and *scriptura* illustrates two contrasting approaches to readings and variant readings. Usually, a reading becomes a variant only when compared to another reading (Froger 80); variant also implies a deviation from a norm, one version of the text which may be chosen at random (Colwell and Tune 253).[4] On the other hand, a variant can be one among multiple possible alternatives, in a place where at least two witnesses disagree as to what the text is. Consequently, Colwell and Tune decided to refer not to variants, but to *variation-units*. This approach is shared by genetic criticism, which reject the existence of an invariant text, against which variant readings are compared (Biasi). In the twentieth century, formal definitions of reading can be found for instance in editing manuals, dictionaries or lexicons. Stussi defines a reading as "a passage from a transmitted text as it appears in a given witness"(Stussi 89).[5] A more precise definition of a reading is given by Froger, while describing one of the first examples of collation software:

> The form or content of the text in a given place is a *reading*, that is to say what we read at this location. Any manuscript, for instance the original, can

---

[2] "*Nam quomodo videri possum emendare Sallustium, qui, incertum est, an sic scriptum reliquerit, ut me tu ais emendare voluisse? Ego tantum* ***ex varia scriptura****, quid mihi satis videatur, pronuncio. At cur praeponis, inquies, illam scripturam, quae in paucioribus codicibus est? Praepono, non ut Sallustius emendem, sed ut admoneam sequendum, quod plurimorum confirmat authoritas.*" (Valla 263). The discussion can be found in Valla's *Antidoti in Pogium*, book I, in the section on Sallustius.

[3] *"non meam, sed variam lectionem accipies illius versus in primo carmine Catulli"* (Avanzi a5v).

[4] Colwell and Tune explain that the "norm" against which variant readings are compared may be different depending on editors: "So what is commonly done in practice? Some particular text is chosen—often at random—for the norm. Either we use a printed text such as the Textus Receptus, sometimes an edition by Tischendorf, Westcott-Hort, or Nestle; or, we may use the text of a particular MS whose textual affinities are already known, e.g., Vaticanus or Alexandrinus" (Colwell and Tune 253).

[5] "Con lezione di un determinato testimone si designa un passo del testo tramandato così come compare in tale testimone" (Stussi 89).

> be considered regarding its content as a collection or set of readings, which are the text elements at various levels: chapter, paragraph, sentence, word, syllable, letter, and even punctuation or accents (Froger 9).[6]

This definition adds more precision: a reading is a textual element ('what is read'), and it can be of various scope, from the smallest punctuation marks to whole chapters. How can these definitions of a reading lead to a first example of a reading model?

2 Modelling a reading

The purpose of data modelling in the Humanities is to describe and structure information about real-world or digital objects in a formal way, so that this information becomes computable (Flanders and Jannidis 229–30) and so that it can be manipulated and queried with the help of a computer in order to answer questions. Ultimately, the purpose of modelling readings is to help determine if two given readings may be considered variant readings in a specific context. Flanders and Jannidis (234) suggest modelling textual variants in a scholarly edition by classifying variants according to some scheme, such as accidental versus substantial, or orthographical versus lexical, which corresponds to a consensus within the community.

As we have seen, however, variants can represent something very different depending on the perspective (stemmatics, linguistics, etc.) and textual traditions (oral, medieval, early printing, and so on); therefore, readings need to be modelled independently of their function in textual criticism, but with enough information to decide what is a variant in those contexts. It may be helpful to consider the distinction between readings and variants in the framework of Sahle's wheel of text model (Sahle 45–49). Readings can be considered as a part of the text as Document (TextD), whereas variants are part of the text as Version (TextF). The text as Version is further divided into subcategories, such as TextK, a canonical representation of the text which aims at identifying the best (true) text. With this framework in mind, the characterization of readings as authentic or corrupt does not make a good model for readings, since it represents rather variants than readings. Therefore, the more recent definitions of readings may provide a better starting point to the model than the true/false distinction previously applied to readings. Models are simplified representations of an object of study, a selection of features among all available (Pierazzo 44–45). From the overview of the term reading provided in the previous section, in particular the

---

[6] "La forme ou teneur du texte en un lieu donné est une «leçon», c'est-à-dire ce qu'on lit à cet endroit. Un manuscrit quelconque, par exemple l'original, peut donc être considéré, quant à sa teneur, comme une collection ou un ensemble de leçons, qui sont les éléments du texte à différentes échelles: celle du chapitre, du paragraphe, de la phrase, du mot, de la syllabe, de la lettre, et même du signe de ponctuation ou des accents" (Froger 9).

definition of Froger and Stussi, features which apply to a reading can be inferred, namely that a reading:

- conveys textual content;
- has a precise location in the text (also referred to as *locus*);
- can occur at any level of the text, and thus have various sizes;
- is transmitted by a witness.

2.1 Issues

These features need to be discussed in more detail. For instance, is it too restrictive to limit a reading to textual content? What about decorations, mathematical diagrams and other non-textual elements? Historians of Greek, Arabic or Egyptian mathematics have acknowledged the need to collate and critically edit mathematical diagrams instead of simply providing corrected figures to fit modern standards. Raynaud created a stemma for the *Epistle on the Shape of the Eclipse* by Ibn al-Haytham, a mathematical treatise from the eleventh century, using the mathematical diagrams present in the text. In order to collate diagrams and apply Lachmann's method of shared errors, Raynaud had to select "characters" from the diagrams, which could be regarded as an equivalent for readings. This suggests that it is possible to define and model readings for mathematical diagrams. It would be different from textual readings, but as important for the comparison of versions from traditions of mathematical texts. Other types of content could include—and are not limited to—visual content such as decorations, illuminations, or artist's sketches (see the contribution of Martina Scholger in this volume, on comparing the sketches of the Austrian artist Hartmut Skerbisch). Musical compositions need as well to be collated and critically edited, however musical readings and variants are quite different from textual readings and variants: for instance, pitch and metrical values are significant features of a musical note to compare (Broude).

Let us focus here on readings as textual content. Other issues arise with gaps and lacunae: can the absence of text, such as an omission, a so-called *lacuna*, be considered a reading as well? It would seem that the absence of text is by definition not a reading. It cannot be read in the witness, even if it can often be defined by the other features listed above (the size of the missing text may be difficult to evaluate in some cases). However, a missing reading may be significant for the manuscript tradition: since a missing passage is difficult to restore by conjecture, a lacuna shared by several witnesses can often be used as a significant error that indicate a relationship between those witnesses (Reynolds and Wilson 213). A lacuna that helps in grouping manuscripts and building the stemma therefore needs to appear in the collation. How should the absence of text be modelled? As a special kind of reading, or separately? In this model, lacunae were included as readings without any content.

Conjectures—reconstructed readings proposed by scholars which are not present in any witness—seem to qualify as readings according to the features listed above. However, one may ask if conjectures are indeed transmitted by a witness. Conjectures are obviously constituted of textual content of a certain size, meant to be read at a certain location; can they nevertheless be considered to be transmitted by a witness when they are published in a scholarly article instead of an edition? According to Greetham, a conjecture "is involved only when an editor reconstructs or creates a reading which is not extant in any of the witnesses" (352). A conjecture is thus a new reading, with no prior witness evidence, but with an established origin that can be traced to a particular scholar or scribe. In this sense conjectures are considered as part of the reading model.

The location of a reading in the text is not as easy to formulate as it seems. It would not be enough, for instance, to number each word, since the count would then be different for every witness. Even a reference system such as the canonical citations for classical texts can have limitations, when it comes to precision at the word level. Citations such as Pliny nat. 11.4.11 or Vergil ecl. 10.69 refer respectively to the *Natural History* of Pliny the Elder, Book 11, Chapter 4, paragraph 11, or Vergil's *Eclogues* 10, verse 69. The minimal text unit here is the paragraph or the verse, not the word, and at some point in the text, there will be chapters or verses with different word numbers. The location in the text can only be accurately expressed after collation has happened and readings have been aligned with each other. Canonical citations have been formalized in digital formats such as DET (Robinson) or the Canonical Text Services (CTS) Data Model (Crane et al.).

Text can be seen as both a conceptual (immaterial) sequence of words and punctuation from which a reader derives meaning and as a material sequence of marks on a document. Readings are also made of marks recorded on a physical document, besides being part of the immaterial text, thus a reading has both a location in the text and a location in the document where it appears. The document location may be rendered with varying degrees of precision: for instance with folio or page number of the witness in which it appears, with an additional line number, or with a very precise set of coordinates for a two dimensional surface on the page.[7] Finally, it is worth asking if different levels of reading (letters, words, sentences and so on) call for different models and how those levels relate to other existing models. For example, how would the letter level relate to the model used by the DigiPal framework Stokes uses to describe letters from a palaeographical point of view? How would the sentence level relate to the treebank model (Haug) used to annotate textual corpora? How would the different levels be linked together, if the intent of the scholar is to collate at different

[7] See, for instance, the TEI P5 Guidelines chapter 11 for representation of primary sources, in particular section 11.1 on digital facsimiles www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PHFAX. Accessed 30 Sept. 2017.

levels? Monella, for instance, decided to collate a text from the Latin Anthology at three different levels, which are called graphical (letters, punctuation), alphabetic (the abstract representation of a letter in a particular alphabet) and linguistic (word) levels.

The different levels may certainly be characterized by additional features of their own. Readings at the word level have a specific spelling, and may be abbreviated. Readings at the word level may also have morphological features, such as lemma or part-of-speech properties or even a phonetic transcription. These linguistic annotations could be useful when comparing readings during collation. For instance, words that do not share gender, number, case and lemma could be considered variants. In the case of oral sources, a different pronunciation may be considered a variant. Layout could also be significant in some contexts: the same word written in bold, or italics or in colour could signal a variation. For instance, Caton argues that a transcription loses information when a word originally written in italics, to denote emphasis, is transcribed into Roman font. At the line level in poetry, metrical patterns would be an important feature. At the sentence level, syntactic information about the subject, object, verb and other elements of the sentence may be an important feature. This information could be particularly interesting for the comparison of versions translated in a different language from the original. In principle, the comparison happens always with readings at the same level: letters are not compared to words, or words to paragraphs. It is worth noting, however, that even if the word level is used during collation, it may be that in the result, words will be grouped together to form a new reading at a different level than the word level (a variation unit that falls between the word and sentence levels). Considering the sentences from the fictive witnesses in figure 1, the groups of words *star shines*, *sun shines*, and *stars shone* may be considered as one reading only, for the purpose of studying the collation results. When there are many variations close to each other, it may be difficult to decide how to group words into readings, if they should be grouped at all, and the readings may be different according to different editors. One could decide to group words instead as *bright star*, *sun* and *bright stars*, with the verb as a separate reading.

2.2 Model

In summary, the model could be expressed as in figure 3: readings can either have content or not. In both cases, a reading has the general features outlined above, such as the witness in which it is found, a position both in the text of the witness and the document of the witness, or a level of precision such as the word level. When the content is present, it can be textual content or another type of content such as diagrams or illustrations. The textual content has a second layer offeatures: syntax, morphology, phonetic, layout, and so on. Depending on the level of the textual content, features may differ. At the sentence level, it is possible to describe the relationships

between words or group of words: *The bright star* is the subject of the verb phrase *shines upon you*, a relationship which is more difficult to represent at a word level. The other types of content would have their own features, such as the characters in diagrams described by Raynaud.

On the other hand, readings without content cannot be described with those additional features. There are other concerns regarding an absence of content, or lacunae. First, there are different reasons behind the presence of a lacuna. The missing text could have been present in the manuscript but is no longer readable by scholars, due to damage or missing pages. In other cases, the copyist marked a lacuna explicitly, with a series of dots for instance, because the text was already missing in the witness serving as the exemplar. The scribe may also have left a blank space to be filled later, and which was never completed. In medieval manuscripts, this would happen easily for materials such as titles, initials or coloured text, which were added later often by a different person than the copyist of the main text. In addition, Dillen has demonstrated the importance of distinguishing between several types of lacunae in Beckett's draft manuscripts, such as authorial lacunae as opposed to editorial ones. Lastly, the lacuna may not be perceptible, unless the witnesses are collated. The collation result could then expose in a witness the absence of a reading which was present in at least one other witness. This kind of lacuna does not belong to the reading model, but only to the variant model: a variant arises either if two readings are considered different, or if a reading is compared against an absence of a reading. In figure 1, the absence of *bright* in witness B would have gone unnoticed unless exposed by the collation against the readings in sentences A and C. The reading may be absent because the scribe did not copy it, whether voluntarily or not, or because it was absent altogether from the exemplar. It is then important to distinguish between the reasons behind a lacuna: is the text present but no longer accessible? Is there a mark indicating that the text was already illegible to the copyist? Or is there no evidence? Even if the text is absent from every witness, the presence of a lacuna can be indicated by inconsistencies in the meaning, for metrical or grammatical reasons, or by incomplete content (such as a missing plural "s").

Given two or more readings at a place of variation, the comparison of the reading's features could help to identify in what aspect the readings differ. This comparison could then lead to a decision regarding which perspective those readings become variant readings of. Let us consider pairs of readings from the sentences in figure 1: comparing the features of *stars* and *star* would show a difference in number, plural and singular, but the lemma would indicate that they represent the same word. It would thus be a grammatical difference. The readings *sun* and *star* have a different lemma, and therefore represent a lexical difference. Two words which share all features (lemma, part of speech and so on) and show no other difference than their original written form would represent an orthographical difference, or graphical difference for

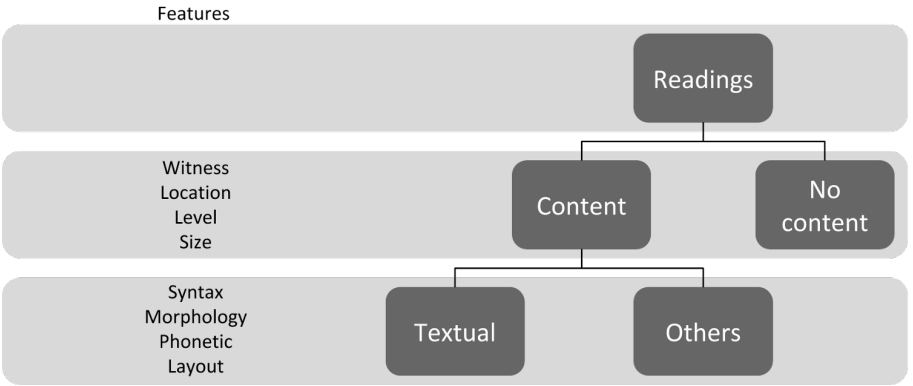


Figure 3: Model for readings.

languages which have no standardized orthography. In different scholarly contexts, the features of readings could be used todefine criteria which are then applied to isolate the relevant variant readings.[8] First, if all differences are considered variants, then readings which display any difference among their features will be considered variants. On the other hand, since orthographical differences are often not considered variants while editing a text (Reynolds and Wilson; Love), the distinction between non-orthographical or orthographical differences allows the editor to select the set of readings which represent grammatical or lexical differences and ignore spelling variants. Finally, linguists would be able to select only spelling variants, particularly significant for the study of language evolution (Vierros and Henriksson). These three contexts will be further examined in Section 4 below, using a practical example. The next section will first deal with the representation of a reading in digital format.

3 Digital representation: from reading to token

To translate the concept of a reading, as defined by centuries of textual scholarship, into digital representation, it seems there is already a counterpart in computational linguistic terminology: the token. Tokens are commonly used for lexical analysis in computer science, as a sequence of characters with an identified *meaning* is converted into a token (see, for instance, Grefenstette and Tapanainen). If manual collation is the comparison of readings, computer-supported collation is the comparison of tokens. Computer-supported collation is the application of computing methods to the

[8] These criteria would not necessarily be applied at the time of recording variants, but also after variants are recorded, to identify only the variants relevant to a specific context.

comparison of textual witnesses: instead of comparing manually the existing versions of a text, digital transcriptions are collated with the help of an alignment algorithm. Juxta and CollateX are two of the most well-known collation tools available, and were both conceived according to the Gothenburg model of collation.

The Gothenburg model was devised in 2009 in order to define computer-supported collation. It divides the process of collation into four successive tasks (Dekkeret al.). The first of these tasks is to split the entire text of each witness into smaller units, called tokens, to be compared. The other tasks include alignment of those tokens (the actual collation process), analysis, and output of the collation results. The parallel between Froger's reading definition (see above) and a token is clear. In the Gothenburg model, a text is divided into a list of tokens which are textual units (a sequence of characters) at a chosen level. This is also how Froger describes a text, as a collection of readings, which are made of the text's content taken at a particular level. As such, the tokens share the same features as readings: the textual content of a witness, with a precise location in the text determined by its position in the full list of tokens, and at a specific level.

According to Dekker et al. (4), a token is a textual unit at "any level of granularity, for instance, on the level of syllables, words, lines, phrases, verses, paragraphs, text nodes in a normalized XML DOM instance, or any other unit suitable to the texts at hand." The CollateX documentation more explicitly considers a token as a textual unit that ideally carries meaning, thus above the character level.[9] At letter level, phenomena such as transposition are much more frequent and reduce the efficiency of the alignment algorithm. For this reason, collation is preferably performed at a higher level, rather than at character level. However useful for the collation process, this restriction does not apply in palaeography where letters are the comparison units. Projects such as Monella's also require analysis at character level. From a theoretical and modelling perspective, it is thus necessary not to make assumption about the meaning of a token. The transcription model of Huitfeldt, Marcoux, and Sperberg-McQueen provides a more adapted description for a token, since they do not make a distinction between tokens as characters, as words, or as other levels.[10]

In digital format, the most basic form of a token is a simple string of characters, a linear sequence of one or more symbols representing letters, but with no linguistic in-

[9] See the CollateX documentation: collatex.net/doc/#tokenization. Accessed 27 Oct. 2016.

[10] "A mark is a perceptible feature of a document (normally something visible, e.g. a line in ink). Marks may be identified as tokens in so far as they are instances of types, and collections of marks may be identified as sequences of tokens in so far as they are instances of sequences of types. In other words, a mark is a token if, but only if, it is understood as instantiating a type. The distinction among marks, tokens, and types may be applied at various levels: letters, words, sentences, and texts" (Huitfeldt, Marcoux, and Sperberg-McQueen 297). The transcription model is "agnostic about whether the types (and tokens) it is concerned with are those at the character level or those at the level of words and lexical items" (Huitfeldt, Marcoux, and Sperberg-McQueen 298).

terpretation attached to them. Nevertheless, collation tools usually offer to normalize tokens in order to minimize what is perceived as insignificant variation: typically, normalization permits the removal of upper case, punctuation or other aspects (such as, for instance, hyphenation or line breaks in Juxta, white space characters in CollateX) from the tokens that will be compared, so that these would not be considered differences: *the* and *The* would be treated as the same word for the purpose of aligning the versions together. However, if this normalized form is not explicitly included in the token, it will not be available in the results of the collation. For example, in the case when accidental differences are not significant, the pair of readings *sun*/*sunne* and *world*/*worlde* may be considered as irrelevant differences and thus should be ignored when searching for semantic variants. However, given only the string of characters it is impossible to discriminate between a significant variant such as *shines*/*shone* and the orthographical variants such as *world*/*worlde*. On the other hand, if the reading *worlde* also includes a normalized form *world*, it is then possible to compare the normalized form of *worlde* and decide that it is equivalent to the reading *world*. As a consequence, it could be extremely difficult to distinguish between orthographical or non-orthographical differences without normalized forms, when analysing collation results.

3.1 Token format in CollateX

CollateX makes it possible to distinguish between the original token and a normalized form provided by the user thanks to an input format in JSON, a lightweight data-interchange format.[11] The structure of CollateX's JSON input is described in the CollateX Documentation (2013). Tokens can therefore be represented as JSON objects with various properties, such as:

- t: the textual content in its original form.
- n: a normalized form of the same textual content.

The normalized form is used to align the texts as accurately as possible, while the original content is still available should it be needed by the user when analysing the results; the JSON format of CollateX is thus a very effective way to represent readings involving textual content. However, the absence of content is problematic, since a token must always have at least a property *t* with a positive value. As a result, it is not possible to collate empty tokens, which is a limitation since lacunae are considered readings in this model and need to be represented as tokens as well. So far, I have represented lacunae present in the text due to damage, or explicitly marked by the copyist, as tokens with the textual content *t* as "…", and the normalized form *n* as

[11] See www.json.org. Accessed 10 Mar. 2017.

"lacuna", a combination of content that does not appear elsewhere in the witnesses and therefore cannot be confused with another reading. Lacunae which are revealed by the collation, because a portion of text was omitted by a scribe, are not represented by a token. Instead, CollateX inserts empty tokens in the collation to compensate for the absence of text (Dekker et al.).

As CollateX is used in other projects, their encoding choices may provide further ideas about the representation of readings as tokens. As an example, the Collation Editor, a tool prepared for the collation of the Greek New Testament with CollateX, provides a description of the token's properties online.[12] The Collation Editor provides two layers of normalization and regularization: the original token is normalized in a first step into *t*, with operations such as setting the words in lower case. Then, the token *t* may be regularized again into *n* according to rules defined by the user, which are provided through a *rule_match* feature.

Besides *t* and *n*, any additional properties can be provided to the token object, but will be ignored during collation. Nevertheless, these additional properties would still be available in the results for further processingsuch as visualization. For tokens at the word level, such properties could also include:

Identification. A way to identify and locate the token in the document where it appears, with a reference to page and line numbers for instance. The location may also help to situate the token in the text (with a reference system, such as canonical citations for classical texts mentioned above). A unique identifier could also serve to link the collation result to the transcription, where other properties of the token are encoded and could be retrieved. The Collation Editor includes properties such as *index*, *siglum*, *verse* and *reading* in order to provide identification for each token.

Markup. XML transcriptions of the witnesses are often used within collation software. Since a lot of valuable information is already encoded in the transcriptions, including layout information, several projects have decided to keep the markup in the token properties. It could be exploited during the collation process: for instance, a word marked as bold could be considered as different from the same word in italics. It could also serve to display tokens with more precision. The Beckett Digital Manuscripts project, for instance, displays additions and deletions thanks to this markup property.[13]

Facsimile. A reference to a digital image, for instance in the form of a link, could be helpful to visualize the original reading in context and assess the transcription accuracy (see Nury).

---

[12] The Collation Editor is a tool produced by The Institute for Textual Scholarship and Electronic Editing (ITSEE) at the University of Birmingham. It is an open source tool available on Github: github.com/itsee-birmingham/collation_editor. Accessed 1 Feb. 2017.

[13] See the update from 17 Sept. 2014 here:www.beckettarchive.org/news.jsp. Accessed 31 Oct. 2016.

Linguistic properties. Linguistic properties could be expressed with a standardized format of detailed linguistic annotation, such as part-of-speech, and morphology. Although Crane argues that morpho-syntactic analysis is one major feature of a digital edition, Monella (184) recognizes that the additional workload may be an issue for the encoder. The use of semi-automatic annotation methods still needs to be explored in further research. Smith and Lindeborg propose to use a "dictionary form" to recognize identical lexical readings, and metrical units to compare the rhythm of Iliadic verses. The use of lemma, synonyms and part-of-speech tagging is also planned to be implemented in collation with the tool iAligner (Yousef and Palladino).

Lacunae. If lacunae are represented as tokens, a description of the lacuna's length and reason (such as damage, or missing pages) could be added. In the Collation Editor, lacunae are not represented as tokens, but are included in the properties of the preceding token: *Gap_after*, a boolean variable set to true, records the presence ofalacuna aftera given token. Another property, *Gap_detail*, gives information about the length of the lacuna.

4 Comparing tokens in different contexts

As described above in Section 2.2, tokens can be compared to find variant readings according to a specific perspective. Three possible situations were taken into account: (a) every difference is a variant, (b) only non-orthographic differences are variants, and (c) only spelling differences are variants. Using the properties *t* and *n* of JSON tokens already make it possible to distinguish variants for these three different contexts. Let us consider again the example of a collated sentence in figure 1. The reading *sunne* was normalized to *sun* and the reading *worlde* was normalized to *world*.

In the first situation, all differences are variant readings. Therefore, in each column, the tokens are compared on the basis of their property *t*: in the first column, all tokens have the same property *t*, *The*, and thus there is no variant. In the second column, the absence of *bright* in witness B is a variant, and so on. When each reading has been examined, the following figure 4 highlights every variant.

In the second scenario, orthographic differences are irrelevant. In order to find the relevant variant readings, the tokens must then be compared on their normalized property *n*, so that orthographic differences appearing in property *t* are ignored. In our example, this means that the last column will not show a variant, because witnesses C and D will have the word *world* as a normalized form: when comparing this normalized form to the tokens in witnesses A and B, there will be no difference. The two tokens show a spelling difference (in property *t*) but are in fact considered the same reading because they share the same property *n*. figure 5 shows non-orthographic variants.

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 4: Every difference is a variant.

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 5: Non-orthographic differences are variants.

Finally, orthographic variants can be isolated when searching for tokens which share the same normalized form *n*, but not the same original form *t*. In our example, there are thus two columns which contain an orthographic variant (see figure 6). The table could then be reduced to a list of orthographic variants only:

1. sun (B) – sunne (D)
2. world (ABC) – worlde (D)

These three simple examples are of course generalizations: in reality, the principles of collation maybe far more complex. For example, spelling differences may be ignored, except in proper nouns (Love 52). In some cases, it may be difficult to distinguish between a spelling difference or a morphological one. In addition, different readers may give diverse interpretations for certain words or sentences, as it is the case with annotated treebanks (Bamman and Crane). Uncertainty and multiple interpretations thus need to be represented as well. However, if the tokens contain more detailed information, it may help to bring more precision when deciding which readings should be considered as variant readings.

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 6: Orthographic differences are variants.

5 Conclusion

Different versions of a text are characterized in part by their variant readings. To represent variant readings in digital format, it may be helpful to precisely define and formalize the concept. What is a variant reading, however, is highly dependent on the tradition in question (oral, medieval, early print, etc.) and on the scholarly perspective on the text (stemmatics, linguistics, and so on, following Sahle's wheel of text for instance). As a result, the set of differences present in a textual tradition are not all considered significant in every situation; variant readings are only a subset of all the differences, and different contexts call for different sets of variant readings, as we have seen in the last section.

A first step in formalizing variant readings may be to model and formalize *readings*, in such a way that later, those readings can be compared efficiently in order to define which readings are considered to be variant readings in a given context. The definitions of the term reading thus provided a series of features which can be used to create a model of a reading. However, those features raised a few issues regarding their content, their position in the text as well as in the document, and their relationship between different levels of reading (from characters to words, sentences, and so on). Following the discussion on these issues, a model was proposed that distinguishes between readings with content or without content. The readings with content can again be divided according to the type of content, such as textual or non-textual. The translation of *readings* to *tokens*, using the CollateX JSON format, showed how the use of a simple normalized form could allow to find different sets of variants in practice, within collation results, according to three different contexts. However, as more information is associated with a reading, it could be possible to define variant readings even more precisely. The aim of the model is to represent readings independently of their function in textual criticism, but with enough information so as to decide when a difference becomes a variant. Considering other sorts of content,

such as mathematical diagrams, images or music, the model is flexible enough to for future extension to incorporate other types of content, such as non-textual readings as well.

Bibliography

Andrews, Tara. "The third way: philology and critical edition in the digital age." *Variants*, vol. 10, 2012, pp. 61–76.

Avanzi, Girolamo. *In Val. Catullum et in Priapeia emendationes.* Tacuinus, 1495.

Bamman, David, and Gregory Crane. "The Ancient Greek and Latin Dependency Treebanks." *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, 2011, pp. 79–89.

Barbrook, Adrian C. et al. "The phylogeny of *The Canterbury Tales*." *Nature*, vol. 394, 1998, p. 839. *Astrophysics Data System*, adsabs.harvard.edu/abs/1998Natur.394..839B. Accessed 25 Oct. 2013.

Beckett, Samuel. *Digital Manuscript Project.* 2016. www.beckettarchive.org. Accessed 12 Nov. 2017.

Broude, Ronald. "When Accidentals are Substantive: Applying Methodologies of Textual Criticism to Scholarly Editions of Music." *Text: Transactions of the Society for Textual Scholarship*, vol. 5, 1991, pp. 105–20.

Caton, Paul. "Lost in Transcription: Types, Tokens, and Modality in Document Representation." *Digital Humanities 2009. Conference Abstracts. University of Maryland, College Park June 22–25, 2009.* Maryland Institute for Technology in the Humanities (MITH), 2009, pp. 80–2.

Cerquiglini, Bernard. *Éloge de la variante: Histoire critique de la philologie.* Seuil, 1989.

CollateX. The Interedition Development Group, 2010-2017. collatex.net. Accessed 17 July 2014.

Colwell, Ernest Cadman, and Ernest W. Tune. "Variant readings: classification and use." *Journal of Biblical Literature*, vol. 83, no. 3, 1964, pp. 253–61. *JSTOR*, www.jstor.org/stable/3264283. Accessed 15 June 2015.

Crane, Gregory. "The Digital Loeb Classical Library - a view from Europe." *Perseus Digital Library Updates*, 2014. sites.tufts.edu/perseusupdates/2014/09/22/the-digital-loeb-classical-library-a-view-from-europe. Accessed 31 Oct. 2016.

Crane, Gregory et al. "Cataloging for a billion word library of Greek and Latin." *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage - DATeCH '14*, pp. 83–8, 2014. *ACM Digital Library*, dl.acm.org/citation.cfm?id=2595188.2595190. Accessed 12 Apr. 2016.

Dane, Joseph A. "The Notion of Variant and the Zen of Collation." *The Myth of Print Culture: Essays on Evidence, Textuality and Bibliographical Method*, edited by Joseph A. Dane, University of Toronto Press, 2003, pp. 88–113.

De Biasi, Pierre-Marc. *La génétique des textes.* Natan, 2000.

Dekker,Ronald, et al. "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project." *Literary and Linguistic Computing*, 2015, vol. 30, no. 3, pp. 452–70. doi.org/10.1093/llc/fqu007. Accessed 28 Oct. 2016.

Dillen, Wout. "'(Hiatus in Ms.)'. Towards a TEI compliant typology of textual lacunae in Samuel Beckett's manuscripts." *Manuscrítica. Revista de crítica genética*, vol. 28, 2015, pp. 65–73.

Epp, Eldon. J. "Towards the Clarification of the Term 'Textual Variant'." *Studies in the Theory and Method of New Testament Textual Criticism*, edited by Gordon D. Fee and Eldon J. Epp, Eerdmans Publishing, 1993, pp. 47–61.

Flanders, Julia, and Fotis Jannidis. "Data Modeling." *A New Companion to Digital Humanities*, edited by Susan Schreibman et al., Blackwell Reference online, 2015, pp. 229–37. www.blackwellreference.com/public/tocnode?id=g9781118680643_chunk_g978111868064318. Accessed 5 Mar. 2017.

Froger, Jacques. *La critique des textes et son automatisation.* Dunod, 1968.

Greetham, David. *Textual Scholarship: An Introduction.* Garland, 1994.

Grefenstette, Gregory and Pasi Tapanainen. "What is a word, what is a sentence? Problems of tokenization." *COMPLEX 1994: 3rd conference on computational lexicography and text research, Budapest, Hungary, 7-10 July*, 1994, pp. 79–87.

Greg, Walter W. "The Rationale of Copy-text." *Studies in Bibliography*, 1950-51, vol. 3, pp. 19–36.

Haug, Dag. "Treebanks in historical linguistic research." *Perspectives on Historical Syntax*, edited by Carlotta Viti, Benjamins, 2015, pp. 188–202.

Havet, Louis. *Manuel de critique verbale appliquée aux text es latins.* Librairie Hachette, 1911. *Internet Archive*, archive.org/details/manueldecritique00haveuoft. Accessed 5 July 2015.

Heikkilä, Tuomas. "The Possibilities and challenges of computer-assisted stemmatology: the example of Vita et miracula s. Symeonis Treverensis." *The Analysis of Ancient and Medieval Texts and Manuscripts: Digital Methods*, edited by Tara Andrews and Caroline Macé, Brepols, 2014, pp. 19–42.

Huitfeldt, Claus, et al. "What is transcription?" *Literary and Linguistic Computing*, vol. 23, no. 3, 2008, pp. 295–310.

Juxta Commons. The Nineteenth-century Scholarship Online (NINES), University of Virginia. www.juxtacommons.org. Accessed 27 Oct. 2016.

Love, Harold. "The Ranking of Variants in the Analysis of Moderately Contaminated Manuscript Traditions." *Studies in Bibliography*, vol. 37, 1984, pp. 39–57. *JSTOR*, www.jstor.org/stable/40371792. Accessed 21 Oct. 2013.

Maas, Paul. *Textual criticism.* Translated by Barbara Flower. Clarendon Press, 1958.

Monella, Paolo. "Many witnesses, many layers: the digital scholarly edition of the *Iudicium coci et pistoris* (Anth.Lat. 199 Riese)." *Digital Humanities: Progetti Italiani Ed Esperienze Di Convergenza Multidisciplinare. Atti Del Convegno Annuale Dell'Associazione Per L'Informatica Umanistica E La Cultura Digitale (AIUCD) Firenze, 13–14 Dicembre 2012*, edited by Fabio Ciotti, Quadernidigilab, 2014, pp. 173–206.

Montanari, Franco. "From Book to Edition: Philology in Ancient Greece." *World Philology*, edited by Sheldon Pollock et al., Harvard University Press, 2015, pp. 25–44.

Nagy, Gregory. "The Homer Multitext." *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia March 26–28, 2010*, edited by Jerome McGann et al., Rice University Press, 2010, pp. 87–112.

Nury, Elisa. "Visualizing Collation Results." *Advances in Digital Scholarly Editing*, edited by

Peter Boot et al., Sidestone Press, 2017, pp. 317–31.

Petti, Anthony E. *English Literary Hands from Chaucer to Dryden.* Edward Arnold, 1977.

Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods.* Ashgate, 2015.

Raynaud, Dominique. "Building the stemma codicum from geometric diagrams." *Archive for History of Exact Sciences*, vol. 68, no. 2, 2014, pp. 207–39.

Reynolds, Leighton D., and Nigel Guy Wilson. *Scribes and scholars. A guide to the transmission of greek and latin literature.* 3rd ed. Clarendon Press, 1991.

Rizzo, Silvia. *Il lessico filologico degli umanisti.* Edizioni di storia e letteratura, 1973.

Robinson, Peter. "Some principles for making collaborative scholarly editions in digital form." *Digital Humanities Quarterly*, vol. 11, no. 2, 2017. www.digitalhumanities.org/dhq/-vol/11/2/000293/000293.html. Accessed 10 Oct. 2017.

Sahle, Patrick. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung.* Books on Demand, 2013. kups.ub.uni-koeln.de/5353. Accessed 3 Nov. 2014.

Salemans, Ben. *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian, way: the case of fourteen text versions of lanseloet van denemerken.* Katholieke Universiteit Nijmegen, 2000.

Smith, David Neel, and Stephanie Lindeborg. "Comparing Digital Scholarly Editions." *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, 2016, pp. 686–7.

Stokes, Peter. "Modeling Medieval Handwriting: A New Approach to Digital Palaeography." *DH2012 Book of Abstracts*, edited by Jan Christoph Meister et al., University of Hamburg 2012, pp. 382–5. www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-medieval-handwriting-a-new-approach-to-digital-palaeography/. Accessed 23 March 2017.

Stussi, Alfredo. *Introduzione agli studi di filologia italiana.* Il Mulino, 1994.

TEI Consortium. "11.1 Digital Facsimiles." *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 3.2.0. Last updated on 10th July 2017. www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PHFAX.

The Etymologies of Isidore of Seville, edited and translated by Stephen A. Barney et al., Cambridge University Press, 2006.

Valla, Lorenzo. *Laurentii Vallae Opera.* Apud Henricum Petrum, 1540.

Vierros, Marja, and Erik Henriksson. "Preprocessing Greek Papyri for Linguistic Annotation." forthcoming, *Archive ouverte HAL*, https://hal.archives-ouvertes.fr/hal-01279493. Accessed 8 Feb. 2017.

Yousef, Tariq, and Chiara Palladino. "iAligner: A tool for syntax-based intra-language text alignment." *Fifth AIUCD Annual Conference*, Venice, 2016, pp. 201–5.

Case Studies

Pieces of a Bigger Puzzle: Tracing the Evolution of Artworks and Conceptual Ideas in Artists' Notebooks

Martina Scholger

Abstract

Artist's notes are a rich source for understanding the motivations behind an artwork, but have been largely neglected by both art history researchers and in scholarly editing. Using the digital edition of the notebooks of the Austrian artist Hartmut Skerbisch as a case study, this article discusses the various methodological approaches to versions in different disciplines—(digital) scholarly editing, musicology, and art history—and their transferability to artists' notes. It explores where versions can be found in a single autograph, in contrast to multiple witnesses, and how they can be represented digitally. Special attention is given to the versioning of graphics—prominently used as form of expression in the relevant notebooks—proposing a model for their formal description which makes them more comparable and reveals different versions and, consequently, the artistic development process.

1 Introduction

Ideas do not arise out of nowhere: they are the result of extensive processes of association and thought experiments. Note-taking may seem spontaneous, but notes are the result of a process of learning to write in a way that will communicate with the future reader (Mach 51). Permitting an idea to come to fruition requires a willingness to record it and to keep it as a note (Barthes 153). In a literary context, a note functions as a hinge between the source material and the text version (Van Hulle 53). Equivalent to this, the note of a visual artist can fulfil a *double hinge* function: between source (i.e., notes), artistic concept and the manifestation of the concept.

This paper investigates—using methods borrowed from (digital) scholarly editing, musicology and art history—the various definitions of *versions* in different disciplines, as well as to what extent these ideas can be transferred to artists' notes. Based on this examination, the paper addresses both the current possibilities and shortcomings of formal digital representations of versions in artists' notebooks, giving special attention to the similarities and differences in a series of textual and graphical modifications undertaken over periods of time.

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 27–56.

Background to a case study: Harmut Skerbisch and conceptual art

In the 1960s, a new art movement emerged, originating in the United States of America and in Europe. Through this new movement, the concept, idea, and process of art production moved to the foreground, overshadowing the predominant emphasis on the final art product. Coined as *conceptual* in 1961 by the Fluxus artist Henry Flynt, the theoretical foundation for this transnational movement was provided by the artists Sol LeWitt and Joseph Kosuth, with their programmatic texts in the late 1960s. Set against the context of this paper, one statement from LeWitt's paragraphs on conceptual art and the relevance of the idea and the thought process in the development of artworks, seems particularly applicable:

> If the artist carries through his idea and makes it into visible form, then all the steps in the process are of importance. The idea itself, even if not made visual, is as much a work of art as any finished product. All intervening steps—scribbles, sketches, drawings, failed works, models, studies, thoughts, conversations—are of interest. Those that show the thought process of the artist are sometimes more interesting than the final product (LeWitt 82).

Here, drafts, notes and sketches as components of a larger *meta-artwork* were considered equal to the final executed work of art (in those cases where an object orientation, i.e. a precise aim to progress towards a final object, even existed) or *became* the artwork itself. After this point installations, happenings and performative acts were recognized as new forms of artistic expression. The lack of permanent physical manifestations, as well as the temporary and ephemeral character of these kinds of art require their own form of documentation on the history of their origins.

The following considerations on versions in notes, as precursors of artistic concepts and works of art, will be exemplified by the notebooks of the Austrian visual artist Hartmut Skerbisch (1945–2009). Although Skerbisch cannot be clearly assigned to a specific art movement—his work ranges from conceptual art to media art and object art—his 35 notebooks are without any doubt conceptual by nature. Over a period of almost 40 years, the artist used them for the conception and development of his artistic ideas: his experiments of thought, for forming his general understanding of artistic concepts, and the detailed planning of his executed works of art. Depending on his focus and purpose, Skerbisch expressed himself alternating between textual and graphical form (Scholger, "Assoziationsprozessen auf der Spur" 38).

Notes as versions

In the context of this volume's topic, namely *versioning*, but also considering the process of editing in general, the question arises around whether and how it is possible to actually capture the versions of this specific type of artistic creation process.

Which kinds of versions exist in artists' notebooks, i.e. a single witness, and are these—graphical and textual notes—conceptual modifications and alterations comparable to those more prominently examined in textual criticism with multiple textual witnesses (Shillingsburg, *Scholarly Editing*)? What does the successive development of sketches in Skerbisch's notebooks have in common with the genetic criticism of draft manuscripts of literary texts (Grésillon), the writing process in *Werkstattdokumenten* (workshop documents) of musical works (Appel and Veit), and *Fassungen* (versions) in the context of art production (Hartmann)?

First, however, it is important to establish an understanding of a version in the context of artist's notebooks.[1] According to Peter Shillingsburg and Siegfried Scheibe, a version denotes a certain stage in the life cycle of a work at a certain time (Shillingsburg, *Scholarly Editing* 47*;* Scheibe 207). While their definitions mainly refer to (literary) texts, Bodo Plachta explicitly includes works of art in his definition of a version as a completed or unfinished execution of a work of art, which differs from another execution (136). Daniel Ferrer states that genetic *variants* can be treated as interpretations of earlier versions, whereas a *variation* manifests implicit aspects of the original form (Ferrer, "Genetic Criticism" 62).

> Typically, one speaks of *variants*, when there is a choice between elements regarded as equivalent, and of *variation* when the similar but different elements are juxtaposed in space or in time (Ferrer, "Variant and Variation" 35).

In a visually oriented context—in contrast to a textual one—it is difficult to identify where the *original form* is, since that term could be assigned to any record of the initial idea, the first recognizable note or conceptual drawing, or even the first manifestation of an artistic concept.

The paper starts with an examination of notebooks as a unique genre, a meta-artwork, and a medium for capturing fleeting thoughts and maturing ideas, discussing their peculiarities and their value in reconstructing the development of specific artistic ideas over time and drawing on the exemplary corpus. It will then propose two hypotheses regarding the formalisation of the creation process, essentially combining methods from (digital) scholarly editing, musicology, and art history and practice to propose a working definition of versions in the context of Skerbisch's notebooks. Following a discussion on the practices for formalisation of textual notes, the paper will then move towards a thorough examination of how an equally-rich and precise formalisation of graphical notes can reveal the genesis of an artistic expression, and propose a model suitable for that task. In conclusion, the paper will discuss the benefits of applying such a formalisation in the revelation and chronological placement of versions throughout a notebook corpus.

---

[1] In her contribution to this volume, Elisa Nury elaborates on the terminological differentiation of *variant* and *version* in textual criticism in great detail.

2 Notebooks, an artist's warehouse of ideas

Artists' notebooks allow a view behind the scenes and are a valuable resource for grasping the story and creation process behind artistic activities. The significance and value of artists' notebooks for the examination of ideas and concepts at a specific time in the life cycle of an artwork will be evaluated through a digital scholarly edition (Scholger, *Die Notizbücher*) of the notebooks of Skerbisch, which the artist kept from the summer of 1968 to March 2008, just one year prior to his death. Around two thirds of the notebooks are textual notes, and one third are sketches and formulas, which mostly refer to Skerbisch's artistic work. Only a few entries throughout the corpus deal with personal issues and because of this were documented—but omitted—in the digital representation of the edition, to respect the personal rights of the author and others involved.

Notebooks provide a very intimate view into the author's studio (Radecke, "Notizbuch-Editionen" 27). They contain immediate, unfiltered, and spontaneous thoughts and inspirations, which are collected in a warehouse of ideas for later use. William Somerset Maugham wrote in his preface to *A Writer's Notebook*: "I meant my notebooks to be a storehouse of materials for future use and nothing else" (xiv). Indeed, the notebooks of Skerbisch (figure 1 shows some representative sample pages) seem to be one large collection which, when considered as a unit, provide a macro-perspective on the basic concepts and associative processes of the artist. Because of this, they can be regarded as a meta-artwork accompanying his artistic work. The fragmentary, unstructured and non-sequential textual and graphical notes were not intended for the public.

The texts switch between unrestrained, spontaneous notes on the one hand, and structured, sophisticated records on the other. Some of the entries are accurately dated, while others can only be placed by referencing them to individual works of art or events. Skerbisch cared little for punctuation and orthography, often merely listing seemingly unrelated catchwords.

Besides text, the notebooks contain graphical components such as sketches, constructional drawings and diagrams, which in this context carry at least the same level of complexity and significance as the text itself. Graphics are used to explain complex facts, such as the detailed construction of installations and objects' details from different viewpoints, which could only be captured through their visual components and which cannot be expressed by text. The converse also applies, since not every situation can be represented in images. In other cases, a combination of both text and graphics is needed. In this case, these are inseparable and comparable in terms of expressiveness.

Furthermore, the entries in the notebooks contain innumerable references to other entries within the notebook corpus, to artworks by the artist, to external works from

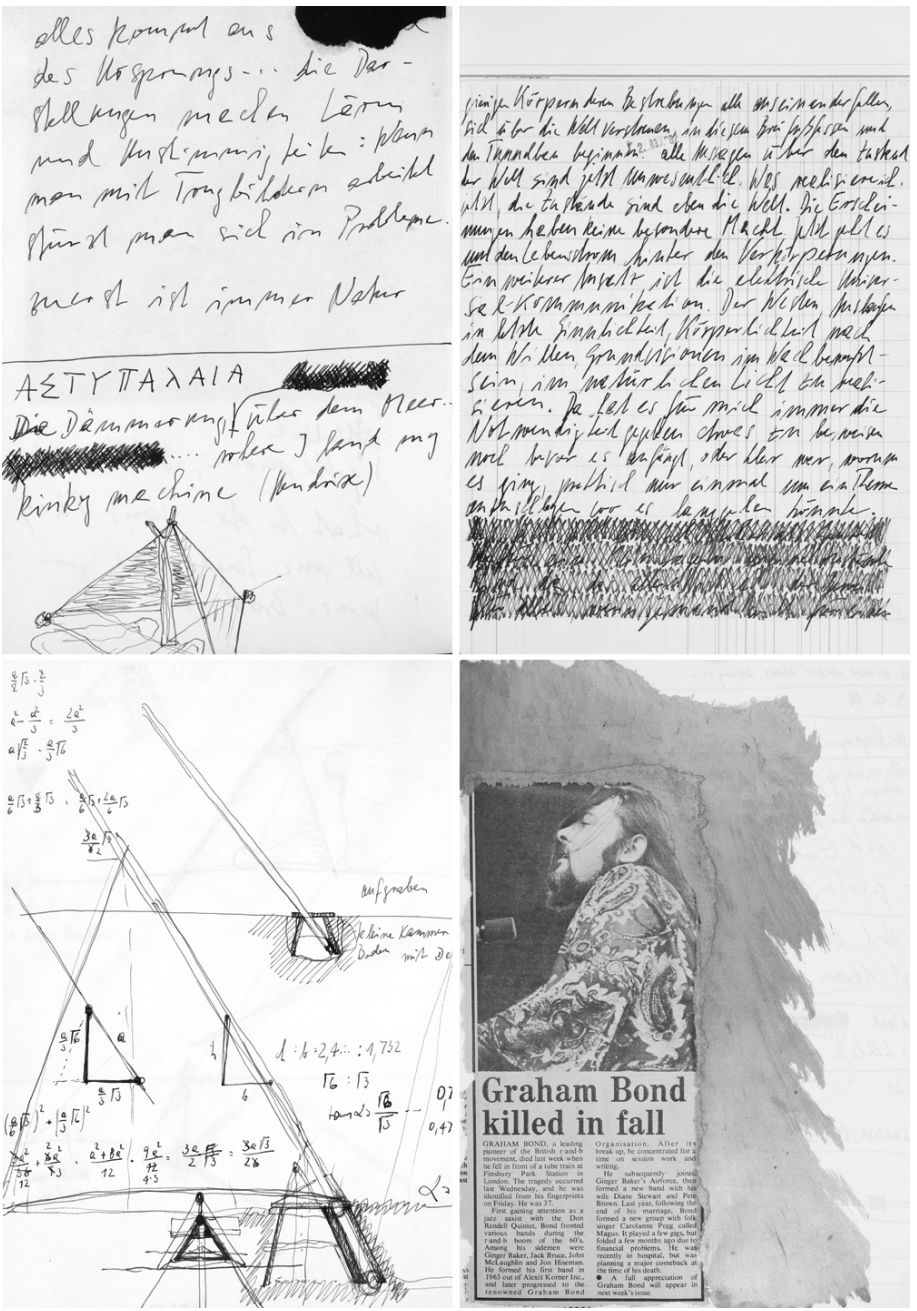


Figure 1: Sample pages from the notebooks.

literature (e.g. James Joyce, Franz Kafka), and to music (e.g. The Rolling Stones, Jimi Hendrix), as illustrated in figure 2. The significance of external influences on Skerbisch's work was mentioned by the artist himself in an interview, when he stated: "I accept the point of view that my work is a commentary" (Fenz 123). The resulting network of references, where entries interconnect through their shared concepts, is particularly appealing to a digital representation, where every single entry can be contextualized within its broader meaning. As will be shown next, the shared ideas and cross references can be understood as a graph, which represents the relationships between objects, their different branches, and (therefore) multiple versions.

3 Tracing the evolution of motifs

The train of thought throughout the notebooks is non-linear and repetitive. The artist engaged with the same topics and ideas several times. To bring these individual traces together and contextualize them against additional material, the genetic and semantically-enriched digital edition of Skerbisch's notebooks focuses on the challenges of tracing how a specific idea evolved and changed over time. The notebooks are his autograph; no additional copies exist and they were primarily used for conceptualizing artistic ideas, which means that the constitution of the text plays a subordinate role to the *unfolding of an artistic idea* (Scholger, "Assoziationsprozessen auf der Spur" 257). Additionally, they ultimately lead to manifestations of these ideas in the form of artworks, performances, and exhibitions, which in Skerbisch's case are not a terminal point, but just one way marker in a bigger conceptual process. As shown in figure 2, this results in a network of direct and indirect relationships between:

a) individual notebook entries consisting of text and graphics;
b) notes and work manifestations;
c) notes and external references to literature, music, persons, and art; and
d) notes and generic intellectual concepts which the artist reflects on.

To handle this aspect adequately in a digital scholarly edition, a semantic layer is needed in addition to the digital representation of the notebooks' contents: drawing on a transcription of the notebooks using the standard of the Text Encoding Initiative (TEI), concepts such as persons, books, or music records are described using the Resource Description Framework (RDF) and linked to established authority files, in turn allowing them to be queried with the graph-based query language, SPARQL.

Figure 3 shows an example of different versions and development stages of the same motif, showing distinct views, projections, proportions and details. Faced with such diversity, it is important to filter out the *essence* of a concept and to determine

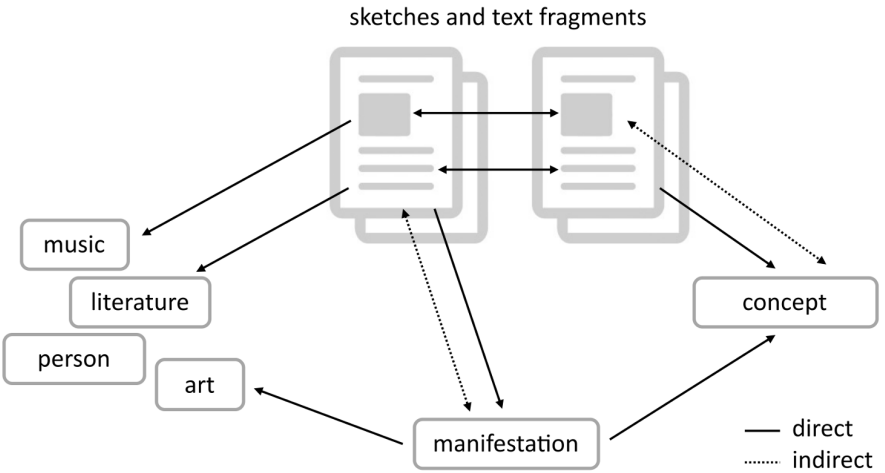


Figure 2: Conceptual model of the artistic circle between idea and manifestation.

what remains *constant*: an approach that applies to different manifestations of both a textual and graphical nature. It should be stressed that this development is not necessarily sequential, but an iterative process.

One striking example for the development of the same motif on the same conceptual level is the intensive work on archaic life processes, making use of various tent and box constructions, in reference to the studies of the architect Predrag Ristić who was interested in the mesolithic settlement culture of Lepenski Vir in Serbia, where the entire civilisation was built around a triangular shape. In his first exhibition at the *poolerie* in 1975, Skerbisch presented two video documentations of his performative acts around the theme of humans and their environment, which prominently displayed the construction of a tent in the first video and a wooden box, the *Kasten*, in the second. A year later, in 1976, he combined the tent and the box in a single work of art, the installation *Erde (Our cubehouse still rocks)*. The connection between these two works can only be determined by their shared concepts, which are evident through the entries in the notebooks connected to the individual works, such as environment, human, land seizure, city foundation and settlement.

Another example where the success of the art installation is inextricably linked to its context is the installation *Zepter und gleißender Stein*, a one-hour exhibition at the Neue Galerie in Graz on December 9, 1977. Here, Skerbisch deconstructed

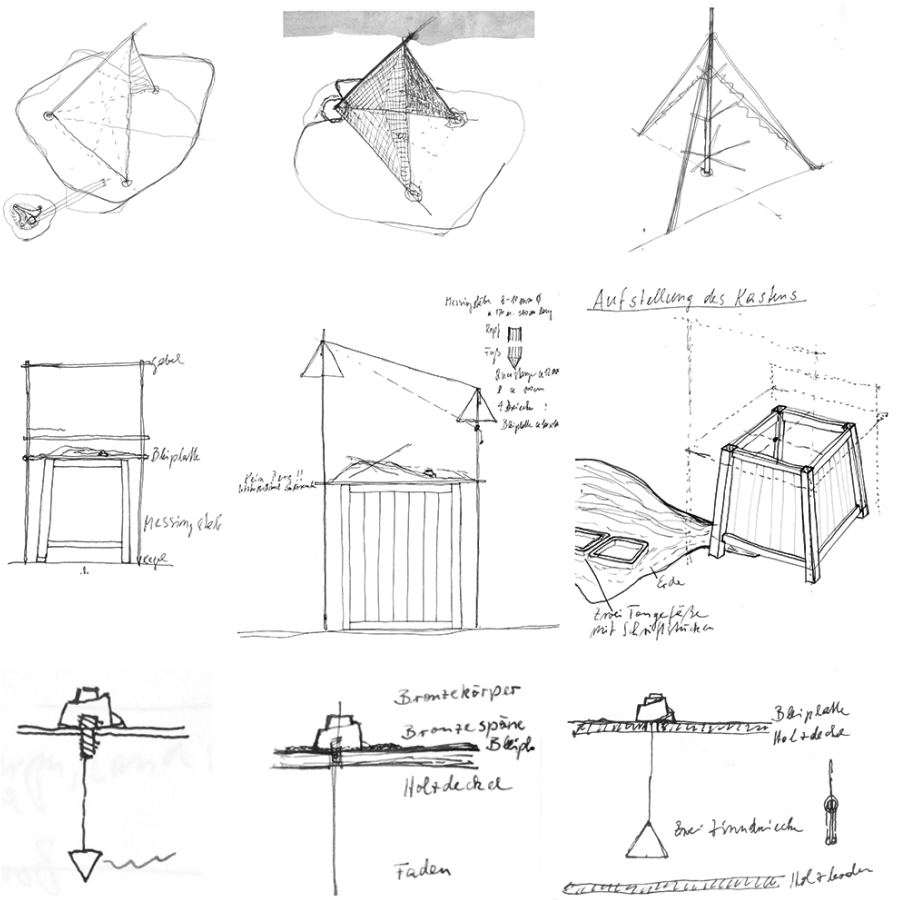


Figure 3: Development stages of the sketches of a tent and a box.

the medium of television and showed its materiality by guiding the visitor through "the interior of the TV system" (Holler-Schuster 160). The visitor was confronted with a wall of 12 television sets displaying a reddish tone, and a camera tube that was presented on red velvet, much like a precious jewel. The work thematized the illusory reality that is fed to the consumer on television by allowing the viewer to walk through the interior of the deconstructed television screen: after exactly one hour, the lights of the installation went out and the exhibition was over. All that remains of this artwork is the written documentation in the notebooks, interview, and photographs taken by both the artist and the exhibition's visitors, which were only published in catalogues, if at all.[2]

A reconstruction of this one-hour installation formed part of the retrospective exhibition of Skerbisch's media works at the Kunsthaus Graz in 2015. Although it gave a rough impression of the situation in 1977, it was decontextualised due to the changed display surroundings—with the installation situated in a large room together with other works of the artist—and the longer duration of the exhibition of more than two months. Neither the spatial experience of a visitor being guided through a TV system, nor the mystification about the abundance of materiality—keeping in mind that in 1977, it was quite extraordinary to have so many screens available for one particular installation—could be recreated. The conceptual considerations of this artwork are reflected in several notebook entries from which two central statements should be highlighted. The first refers to the spectator as an essential *component* of the artwork, Skerbisch states: "diese [sic] Arbeit ist erst fertig, wenn Sie von Besuchern betreten wird" (this work is not finished until it is entered by visitors) (Scholger, *Die Notizbücher*, Notebook 8, 5v). The second statement describes television as the medium that dominates the world: "Das Zepter, durch das die Welt als Erscheinung über allen waltet" (The sceptre through which the world as a phenomenon rules over all) (Scholger, *Die Notizbücher* Notebook 7, 16r). Both of these statements could not be satisfactorily met in the retrospective exhibition from 2015.

These examples demonstrate how the documentation of the artistic processes of conceptualisation, modification and realisation in the notebooks provides invaluable and sometimes even irreplaceable evidence for posterity. A new version of artworks in general, and performative works in particular, is not only dependent on individual components, but also on the original context with regards to time and space, since the experience of the visitors in the original context is hard—if not impossible—to recreate.

---

[2] A similar problem is addressed by Christian Thomas in this volume when he refers to the reconstruction of Humboldt's Kosmos-Lectures, based on reports and written notes from the audience.

4 Methodical approaches to revealing versions

The following sections will investigate editorial methods and artistic genres, which are taken from different disciplines: from (digital) scholarly editing, musicology, and art history. They form the basis for a series of reflections on recording interventions within the texts and the graphics. This should serve to clarify the terminology of version, variant, and variation used in this paper and help to define the research area, without claiming completeness. In order to approach suitable editorial methods, two hypotheses are formulated and will be explored in the following sections: the necessity to consider approaches from both traditional and digital scholarly editing for the identification of versions, and the suitability of artistic practices and methods to define their properties.

Hypothesis 1: The application of terminology from (digital) scholarly editing is necessary for the identification of versions in a notebook.

This section will focus on two text-critical perspectives: a) the *copy-text* theory, mainly concerned with early modern print materials, and b) *genetic criticism*, focusing on (contemporary) draft manuscripts, discussing the benefits of these models for analysing Skerbisch's notebooks.

The copy-text theory is an Anglo-American approach which primarily identifies and removes errors in later witnesses of a text to come as close to the original intention of the author as possible, to "the most authoritative text" (Greg 19). The selected copy-text does not necessarily have to be the earliest, but rather the most reliable. In his essay *The Rationale of Copy-Text* from 1950, Walter Wilson Greg made a crucial distinction between *substantive readings* and *accidentals* of the text, which later was revisited and further developed by Fredson Bowers and Thomas Tanselle, and is now designated as the Greg-Bowser-Tanselle method.

> But here we need to draw a distinction between the significant, or as I shall call them 'substantive', readings of the text, those namely that affect the author's meaning or the essence of his expression, and others, such in general as spelling, punctuation, word-division, and the like, affecting mainly its formal presentation, which may be regarded as the accidents, or as I shall call them 'accidentals', of the text (Greg 21).

The distinction between substantives and accidentals is not only suitable for examining a text, but also graphics in a notebook. An example of this is given in the first three sketches in figure 3: whereas the shift from a cubic to a conic shape of the box must be considered a substantial function of the construction, which is essential

for the installation, the detailed design of individual slats is of secondary importance. However, while the copy-text theory uses this method to highlight the most authoritative text, the focus in the context of the graphics in the notebooks is on the development of a concept and subsequently on artworks manifested outside the written medium.

Another critical approach for taking into account textual variations is the French *critique génétique*, which puts the process of a text's constitution in focus by looking at notes, draft manuscripts, and revisions. Having evolved in France in the 1970s, this post-structuralist method was however originally primarily concerned with the medium of text in general, in particular the process of developing literary works:

> The whole operation [critique génétique] is best described as "genetic" criticism, for it is concerned with literary genesis (even though the term implies a kind of teleology), the whole process of giving birth to the text when finally the obstetrician takes over from the geneticist, to recall a nice distinction Guy Rosa has a good deal of telling fun with (Bowman 628).

In the 1980s Pierre-Marc de Biasi was elaborating on the transmission of genetic criticism to other objects and media beyond literary manuscripts. In particular, the focus of his research was the application of the method to architecture, performances and sculpture (De Biasi, "Pour une approche génétique"). He claims that "the model for genetic analysis that emerges from the study of modern literary manuscripts can, without any possible doubt, be extended to other fields of creation" (De Biasi, "Horizons" 124).

In the context of Skerbisch's notebooks, the question that arises is around how genetic criticism can be transferred from a textual, literary focus to other artistic expressions, where conceptualization and development (writing and drawing) and presentation (performance, installation, sculpture) are carried out in different media, in contrast to literary texts published in written form, where the conceptual process and the product coincide in one medium. Oral narratives and performances of literary texts are excluded from this examination. Again, in the case of his notebooks, the constitution of the text plays a subordinate role to the unfolding of artistic ideas and general concepts.

One proof of concept for the transferability of genetic criticism from literary text to other sources is *Beethovens Werkstatt* (Cox et al.), a research project focusing on the compositional process in Ludwig van Beethoven's (1770–1827) oeuvre, where the critique génétique is joined with digital editing methods, employing the encoding standard of the Music Encoding Initiative (MEI). In musical compositions, there is a differentiation between *closed variants* and *open variants*. While the former is entirely contextualised with the surrounding text at its beginning and end, the latter is not fully connected. Moreover, in closed variants all parts are fully developed. Variants

which were spontaneously created during the composition process are designated as *adhoc*, whereas the correction at a later point after the composition was completed is known as a *revision* (Appel and Veit).

Hypothesis 2: The manuscript in question was composed by an artist. Therefore, an examination of art practices and methods is useful for a digital representation.

Artists record their ideas in a variety of forms, including brief doodles, sketches, and constructional drawings. These graphical representations shed light on the conceptualisation of an artwork and the process of creation. In art theory, there are related terms to express different steps of preliminary stages of an artwork—*sketch*, *study*, *modello* and *preparatory drawing*—which build the basis of an artwork that is often presented in a different medium. Drawing is a means of recording observations and ideas and is defined by its supporting material, by drawing tools and by the formal language used. Fascinated by the pure expression and spontaneity of sketches, Denis Diderot was the first to attribute them authority as an independent artistic form of expression (qtd. in Barasch 127). In the *Dictionary of Art*, a sketch is defined as "rough, preliminary, version of composition" (Turner 817), which is used equally in visual arts, architecture and music. In contrast to the sketch, a study is devoted to individual problems of representation, such as anatomy, perspective, clothing, or movement. A modello or preparatory drawing on the other hand, is a very mature drawing or a three-dimensional model that forms the preliminary stage before the final execution. Models are equally used in painting, sculpture and architecture in order to create a representation that is as accurate as possible (Turner 212–233; Leymarie et al. 40–41).

In the early 20th century, sketches were largely neglected, since artists broke with artistic traditions and started to work directly on the canvas to mimic a spontaneous and immediate situation. In recent years, sketches returned to the spotlight in different (artistic, scientific, technical, etc.) disciplines (Myssok 78) and are regarded as a valuable resource in various cultural mediation and research endeavours. The British Library preserves notebook sketches from Albrecht Dürer's proportion studies, studies on infections during World War One by the bacteriologist Sir Alexander Fleming, and notes on mechanics and architecture from Leonardo da Vinci's *Codex Arundel* in its permanent exhibition. Friederike Fellner investigates the numerous drawings of Franz Kafka in diaries, letters, or on single sheets of paper as part of his literary process. The Zentrum Paul Klee developed a digital edition of Klee's lecture notes during his time at the Bauhaus in Weimar and Dessau (Eggelhöfer and Keller Tschirren). More examples can be found in the digital scholarly editions of Theodor Fontane's notebooks (Radecke, *Theodor Fontane: Notizbücher*) and Vincent Van Gogh's correspondence (Jansen et al.).

In the art historical context, a *Fassung* (version) means a repetition of the same work of art by the artist himself, which shows some modifications to the initial representation. This is referred to as a second or third version. As long as the idea was personally conceived and executed by the artist, it is an *original*. If the artwork is accurately repeated by the artist's studio without any changes, it is referred to as a workshop *replica* or *reproduction*, whereas a repetition made by other artists is known as a *copy* (Hartmann; Rosen 120–121).

A genre of its own which uses repetition of a motif in different versions as an artistic element is *serial imagery*. Paintings such as Claude Monet's *Water Lilies* or LeWitt's *Cubes* are famous representatives of this art form; however, there are also examples from poetry, such as Gertrude Stein's *Sacred Emily*, which has the famous verse "Rose is a rose is a rose is a rose". Skerbisch also produced such series with his *spheres* and *fractals*, a result of his intensive examination of materiality and geometry in the 2000s. Discussing the method of serial imagery, Katharina Sykora describes the differentiation between *constant* and *variable* elements as crucial. Single objects are not solely connected through their subject, but also through their composition. This terminology strongly resembles the Greg-Bowers-Tanselle method of substantive readings and accidental text. The comparison of various views of the reproduced object allows for an exhaustive interpretation of the artefact. In order to distinguish a *series* from a *variation on a theme*, the latter can be comprehended without the total juxtaposition of its contextualized variations, while the former needs the context of its predecessors and successors (Sykora 6).

Sketches, in their various states of expression, are a valuable resource, appealing through their spontaneous execution and proximity to the original idea, and are increasingly being considered in digital editions. Due to the authoritative nature of Skerbisch's notebooks, we are dealing with originals. The distinction between constant and variable in serial imagery is well suited to investigate Skerbisch's notes with regards to defining the essence of individual versions.

Returning to Ferrer's differentiation between variant and variation, we can discern that Skerbisch's notebooks contain *substantive variants* (content-related versions of equivalent elements) and *marginal variants* (minor corrections, that do not alter the content). Therefore, we can speak of *variations*, when the connection between earlier and later versions consists of shared concepts, rather than specific elements of the texts and graphics in question.

5 Identifying versions in notebooks

The previous sections have discussed different theoretical and methodological approaches, dealing mainly with the process of a work coming into being, rather than

the final product. This section will now ask whether (and where) versions can be found in the notebooks. Where is the version in an autograph which has never been copied or published? It is obviously different to the understanding of version discussed in the context of textual criticism of medieval documents, where the text is created from several text witnesses and cleaned up by emendation. Within the Skerbisch corpus, at least two major categories of versions can be distinguished: 1) versions within the notes themselves and 2) versions in relation to the manifested works of art. Furthermore, there are two sub-categories of versions within the notebooks: a) the versions of text and b) the versions of graphics.

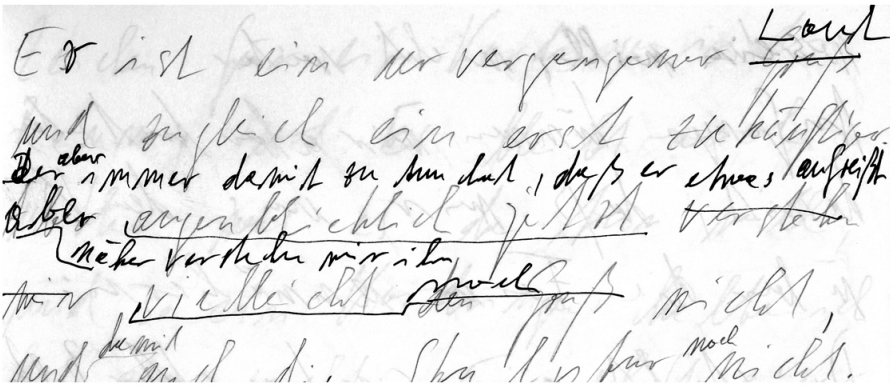
Versions of text

The first type of text version concerns the development of the text on a single document and is created by textual interventions such as additions, deletions, substitutions, transpositions, and alternative readings. Taking the interventions into account, the variant readings reveal different states of text at a specific point in time (Pierazzo 169). For the digital representation of these phenomena, the Text Encoding Initiative (TEI) offers a number of elements and attributes, documented in chapter 11 of the TEI Guidelines (TEI Consortium, "11 Representation of Primary Sources"), which consider the physical document, the encoding of textual interventions and the documentation of the writing process (Burnard et al.).

The notebook entries contain a number of text corrections, some of which are limited to orthographic features, and thus not pertinent to the current discussion. The recording of these interventions becomes much more exciting in those cases in which changes in content take place. Figure 4 shows the facsimile detail alongside the transcription of the initial text state (Level 0) with three revision levels (Levels 1–3), which were identified by the change of meaning, writing instrument, and colour. It documents part of the meticulous planning process for an opening speech at an exhibition.

The second type of text version prominently evident in the notebooks does not take place within a specific text passage on the document, but rather through the mechanism of repetition of particular words, phrases, and sentences throughout the corpus.

Skerbisch used this mechanism to strengthen and sharpen his mind by committing the same or slightly different phrase to paper. It is characteristic that in most instances the artist did not work on the document by setting textual interventions within the text itself, but repeated the phrase without marking the status of the previous mention in any form. An example of this process is the phrase "sie hat angefangen …", which is written 20 times. Three examples from the manuscripts are shown in figure 5, containing a version using abbreviations, a shortened version, and an extended version.



E~~sr~~ ist ein urvergangener ~~Gruß~~ Laut	Level 0
und zugleich ein erst zukünftiger.	
~~d~~Der aber immer damit zu tun hat, daß er etwas aufreißt	*Level 1*
~~Aber~~aber näher verstehen wir ihn augenblicklich jetzt ~~verstehen wir~~ vielleicht ~~den Gruß~~ noch nicht,	*Level 2*
und damit auch die Skulptur noch nicht.	*Level 3*

Figure 4: Conception of an exhibition opening speech in 1992, Notebook 19, 8r (at the top); merging of the four levels of the text (at the bottom).

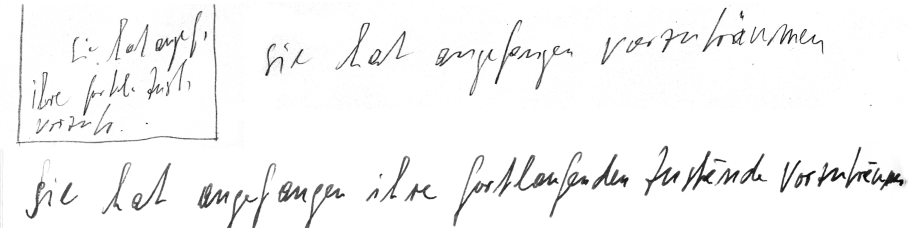


Figure 5: Repetition of the phrase “sie hat angefangen …”.

The first occurrence of the phrase is "sie hat angefangen ihre fortlaufenden Zustände vorzuträumen" (she has started dreaming up her ongoing conditions) (Scholger, *Die Notizbücher* Notebook 8, 11r).

Figure 6 shows a collation of all instances of the phrase. For this purpose, the collation tool CollateX (The Interedition Development Group) has been used for comparing, collating and investigating different versions of the text. Looking at all versions, changes concerning the upper and lower-case writing, the use of punctuation marks, and the completeness of the phrase can be determined. The most noticeable modification relates to the word *Zustände* (conditions), providing three alternative variants: *Ereignisse* (events), *Entfaltung* (development), and *Vorgänge* (processes). The last occurrence of the phrase even formulates a question.

This repetitive approach can be read and interpreted in two ways: a) the latter replaces the previous mention, or b) they are alternatives, equal in their meaning and usage. Tellingly, a clear decision for a final version is not identifiable within the notebook entries by any means, it can, however be found in an external source, an exhibition catalogue from 1978 as a subtitle for a video installation: "sie hat angefangen, ihre fortlaufenden Zustände vorzuträumen" (Künstlerhaus Wien 2).

Both types of textual versions show substantive and marginal variants, building a number of individual versions in time and space. While the substitution from *Gruß* to *Laut* in the first example and the choice between *Zustände*, *Ereignisse*, *Entfaltung*, and *Vorgänge* in the second example must be considered substantive variants, the changes in punctuation and capitalisation can be regarded as marginal variants.

Versioning graphics

Besides text versions, the numerous versions of graphics must be considered. The vast majority of the graphical components in Skerbisch's notebooks are sketches. These have to be examined on two levels: a) their *formal design* and b) their *conceptual meaning*. As far as the former is concerned, alternative versions can be identified on a formal level through the alternation between geometric forms and viewpoints, as well as the positioning of certain elements. The situation becomes more complicated when considering conceptual meaning: Can we still speak of versions when the degree of alteration moves away from the formal level towards a conceptual level? When can we speak of versions of a unique artwork and when of a variation on a theme or even a separate work of art?

Drawing on a similar principle, variant and variation are used to distinguish between two different stages of graphics. These are designated as versions if they cover the same topic and are visually perceived as the same motif with substantive and marginal variants. By contrast, a variation covers the same artistic concept, but varies notably in appearance, i.e. the variables overweigh the constants. The connection is

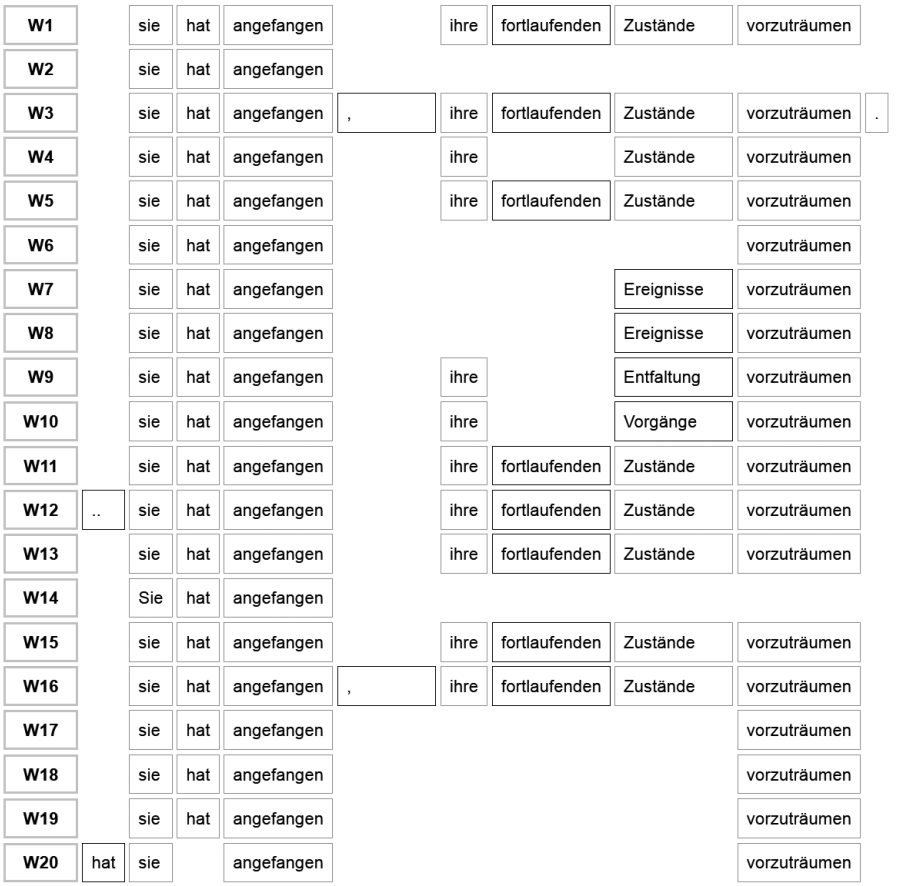


Figure 6: Collation of a phrase repeated 20 times.

not primarily recognisable through the visual appearance any more, but requires a more intensive examination of the artist's entire oeuvre. This applies where an artist experiments with the same object-matter repeatedly, showing the development of an artistic style.

When encoding the text entries, the use of TEI was the evident solution from the outset, since it covers the most common cases for the encoding of primary sources. For the encoding of the graphics, there is a dedicated module of the TEI Guidelines (TEI Consortium, "14 Tables, Formulæ, Graphics and Notated Music") which seems sufficient to begin with. The `<figure>` element is recommended to signal the existence of a graphic in any form: this can be an illustration, a sketch, a photograph, or any other pictorial representation. The `<figure>` element can nest one or more `<graphic>` elements, which refer to the location of the digital image in its `@url` attribute. An additional `<figDesc>` element allows for describing the actual graphic in prose and the `<head>` element can encode a figure caption, if present (TEI Consortium, "14.4 Specific Elements for Graphic Images"). When it comes to text as part of the graphic—let us assume the three box constructions from figure 3—either the structural elements `<p>` for paragraphs, `<ab>` for arbitrary text blocks, or `<label>` for designators, can be used; however, they are not sufficient to convey the full meaning of the content (TEI Consortium, "The element `<p>`"; "The element `<ab>`"; "The element `<label>`").

```
<figure>
  <graphic url="box–01.jpg" />
  <figDesc>Konstruktion eines Holzkastens</figDesc>
  <label>Gabel</label>
  <label>Bleiplatte</label>
  <label>Messingstab</label>
  <label>Kegel</label>
</figure>
```

When focusing on the physical disposition of graphics and their details, the TEI `<sourceDoc>` structure is suitable, as it allows the definition of surfaces `<surface>`, zones `<zone>`, and lines `<line>` with exact coordinates for locating every single graphical component and every single part of text (TEI Consortium, "11.1 Digital Facsimiles").

```
<sourceDoc>
  <surface>
    <zone ulx="288" uly="136" lrx="1500" lry="892">
      <graphic url="box–01.jpg" />
      <zone ulx="932" uly="181" lrx="1086" lry="264">Gabel</zone>
      <zone ulx="946" uly="256" lrx="1138" lry="322">Bleiplatte</zone>
      <zone ulx="940" uly="320" lrx="1156" lry="380">Messingstab</zone>
      <zone ulx="939" uly="380" lrx="1204" lry="438">Kegel</zone>
    </zone>
  </surface>
</sourceDoc>
```

The TEI enables recording the existence of sketches, formulas and other graphical components to be recorded on a text structural `<text>` and topographical level

<sourceDoc>. However, it is lacking when it comes to a detailed formal and content-related description of graphics, the formalisation of complex interactions of text and graphics with entities inside and outside the notebooks, and the recording of alterations within a graphic.

Formal representation of graphics[3]

Graphical elements like sketches are means of artistic expression. Such drawings often convey a message impossible to transmit through text alone: in this case, the graphic becomes the primary carrier. For a comprehensive description of graphical representations, a three-part model is proposed which considers the various (1) graphical components, (2) the textual functions and (3) the interpretations provided by the editor. The first two levels, describing the graphical components and the textual functions—essentially constituting the material record of the source material—are descriptive. The third, interpretational layer functions as an extended commentary contextualising the textual and graphical entries with notebook-internal and -external material related to them.

The proposed model (see figure 7) describes the graphical components of a pictorial representation in the first layer. It declares:

a) the *type* of the graphic representation (e.g. sketch, constructional drawing, doodle),
b) the *projection* (e.g. front view, plan view),
c) the *status* of execution (e.g. total view, detail view),
d) the material of the *information carrier* (e.g. paper, newspaper, photograph),
e) the *drawing instrument* (e.g. pencil, ink pen),
f) *date* or *time span* (to facilitate a chronological order for further investigation on the genesis of the work), and
g) primary graphical *shapes* and *figures* (e.g. triangle, square, cube, tetrahedron).

The second layer records the textual functions, i.e. any explanatory text added to the graphic by the artist. This category includes:

a) *caption*,
b) *description* related to the whole graphic or parts of it, and
c) *label* which designates a specific component of the graphic, sometimes made explicit through a connecting line or clarified through its distinct positioning.

Besides that, the textual content can be of a specific type (e.g. providing information on the *material* or *measurement* proposed in the physical manifestation).

---

[3] Elements of this section overlap with a more detailed technical presentation of the subject forthcoming in the *Journal of the Text Encoding Initiative*, including descriptions of specific elements and the graphic thesaurus developed in the course of the project (Scholger, "Taking Note").

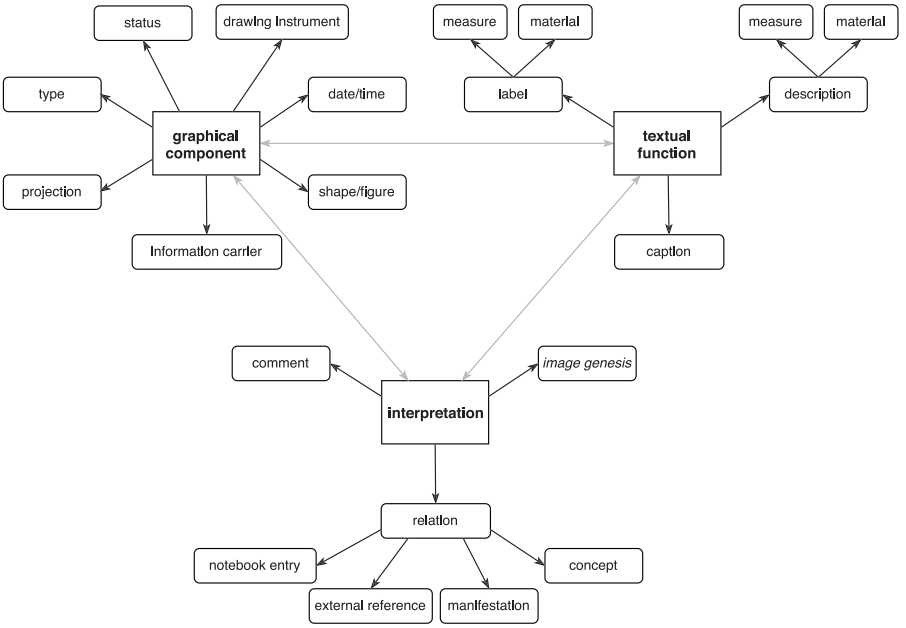


Figure 7: Model for a formal representation of graphics.

The third layer represents the editor's interpretation. It can contain: a) a general *comment*, b) several *relations* to other notebook entries (these can be graphics, but also text); external references to literature, art, music or other preparatory objects (like models or photographs) (*Die Notizbücher* "Register"; "Thesauri"). An implicit part of the third layer is the *image genesis*, which is not explicitly encoded but a result of the annotation process and therefore operative. Versions are generated automatically and can be brought into a sequence to show the development of modifications, comparable to different stages of a text in genetic criticism.

In order to trace the genealogy of the sketches, the interpretational layer needs to record the alterations in comparison to previous entries by pointing to respective locations within the graphics. This is a prerequisite for identifying versions, variants, and variations in the analysis process.

Tables 1–3 show the application of these categories to the sketches of the first two box constructions from figure 3 (middle row). It is easy to spot which graphics can be linked and where changes or alterations occur. Where possible, properties such as shape type, shape and figure, material, notebook entries and external references,

	Graphic 1	Graphic 2
Type	Constructional drawing	
Projection	Back view	Side view
Status	Total view	
Information carrier	Paper	
Drawing instrument	Ink pen, black	
Date	1972-11-15	1972-11-29
Shape/figure	Cube	

Table 1: Instances of the graphical components.

manifestations and intellectual concepts are linked to thesauri/controlled vocabularies from existing authority files or the registers of Skerbisch's artworks and artistic concepts compiled in the course of the digital edition. These same tables reveal that many of the instances remain the same across the representations, especially in the formal description of the graphic and the relations to other entities, but some show distinct differences regarding the textual functions, i.e. Skerbisch's instructions for materiality and measurements of the artworks' envisioned realisation.

By comparing the changes against a formalised model, the alteration of key components is evident. To transfer the demonstrated model into digital structures, the existing methods for encoding graphics in TEI are augmented by a semantic web approach. In the notebooks there are *semantics* communicated by the artist via figures and shapes—a circle stands for a stone slab, a tetrahedron for a tent, etc.—which need to be considered in the digital representation.

Descriptive elements from the TEI encoding are formalised by linking them to specific concepts in decicated thesauri describing artworks, concepts, and graphics expressed in RDF/XML, by employing the `@ana` attribute which "indicates one or more elements containing interpretations of the element on which the `@ana` attribute appears" (TEI Consortium, "att.global.analytic"). This enables not only the representation of the TEI-encoded content, but also the filtering of common factual statements. Georg Vogeler describes this approach to critical representation in digital form as a trinity of image, sign and meaning of content (2015). The thesaurus for formalizing

	Graphic 1	Graphic 2
Caption	–	–
Description	–	Kein Berg‼
Label	Gabel; Bleiplatte; Messingstab; Kegel	Messingstäbe 8-10mm Ø ca. 170 u. 140 cm lang Kopf Fuß Querstange ca 12mm l ca 100cm 4 Dreicke Bleiplatte 100 x 100
Material	Messing, Blei	Messing, Blei
Measurement	–	8-10mm; ca. 170 und 140 cm lang …

Table 2: Instances of the textual functions.

graphical components is a SKOS representation assembled from entities encoded in the TEI document, incorporating concepts from existing authority files, namely the Art & Architecture Thesaurus, in addition to individually-defined concepts where the authority files are lacking, and can be flexibly expanded at any time when new material is processed. It currently contains seven facets: (a) visual works, (b) materials, (c) drafting, drawing and writing equipment, (d) supporting material, (e) geometric figures, (f) views, and (g) interpretation (Scholger, *Thesaurus for Graphics*). The artwork thesaurus represents information on Skerbisch's artworks, like type (installation, photography), date, status (permanent, temporary), and refers to their locations in the case of art in a public space or exhibition venues. The following code example shows a combination of the methods already available in the TEI standard—the recording of the physical dispositions and the intellectual content, and referencing dedicated concepts through the `@ana` attribute. The prefix *art* refers to the artwork thesaurus and the prefix *gt* to the graphic thesaurus.

	Graphic 1	Graphic 2
Comment	Box construction with a forked bracket for mounting a canopy …	Box construction with …
	Relation	
Notebook entry	#TB09-004 (sketch), #TB09-128 (text), #TB10-12 (text)	
External reference	Exhibition catalogue	
Manifestation	#A10076 (Der Kasten)	
Concept	Space, installation, body, environment	
	Image genesis	
Alteration	–	Changing the suspension for the canopy: triangles instead of forks
Deletion	–	The indicated mountain is crossed out

Table 3: Instances of the interpretational layer.

```
<!— section 1: encoding of the physical dispositions in the facsimile structure
     —>

<facsimile>
  <surface xml:id="fol_19v">
    <zone xml:id="F—27" ulx="288" uly="136" lrx="1500" lry="892">
      <graphic url="box.jpg" />
      <zone xml:id="F—27—01—a" ulx="932" uly="181" lrx="1086" lry="264" />
      <zone xml:id="F—27—01—b" points="604,151 596,520 619,524 631,152" />
      <zone xml:id="F—27—02—a" ulx="946" uly="256" lrx="1138" lry="322" />
      <zone xml:id="F—27—02—b" points="614,187 388,511 692,693 746,353"/>

 <!— more <zone> elements —>

    </zone>
  </surface>
</facsimile>
```

```
<!—— section 2: encoding of the sketch in the intellectual text structure ——>

<figure facs="#F—27" ana="art:A10076 gt:1010200 gt:5020000 gt:6010000">
  <figDesc>Konstruktionszeichnung eines Holzkastens</figDesc>
  <figure facs="#F—27—01">
    <label facs="#F—27—01—a" ana="gt:5060000">Gabel</label>
    <graphic url="#F—27—01—b" />
  </figure>
  <figure facs="#F—27—02">
    <label facs="#F—27—02—a">
                <material ana="gt:2110000">Blei</material>platte
        </label>
    <graphic url="#F—27—02—b" />
  </figure>
  <figure facs="#F—27—03">
    <label facs="#F—27—03—a" ana="gt:2020000">
           <material ana="gt:2120000">Messing</material>stab</label>
    <graphic url="#F—27—03—b" />
  </figure>
  <figure facs="#F—27—04">
    <label facs="#F—27—04—a" ana="gt:5070000">Kegel</label>
    <graphic url="#F—27—04—b" />
  </figure>
  <graphic url="#F—27—a" />
</figure>
```

This strategy in searching for similar concepts assigned to graphics is technically implemented via a triple store by using SPARQL queries. Initially, factual information encoded in the TEI document is extracted with XSLT and converted into XML/RDF triples which are then stored in the graph database *Blazegraph* (Systap).

Through this formalisation process, similar graphics can now be extracted from the entire notebook corpus (figure 8), compared and examined for similarities and differences. The resulting XML tree of this SPARQL query extracts a list of images as well as a synopsis of graphical features. The reference to the position in the notebooks gives the temporal information to create a chronology of changes. Drawing on the encoding of the location of graphical details explained above, single components are highlighted in the user interface, supporting the orientation and interpretation process. The comparative juxtaposition, in connection with the temporal dimension of the course of development, makes it possible to reveal specific stages and degrees of completion. With the help of SPARQL queries, all the sketches showing a box can be extracted. The selection can further be restricted to sketches containing both boxes and tents, and so on.

The overall view of the corpus not only gives a picture of the chronological sequence in which the drawings occurred, but also shows the intensity with which the artist has devoted himself to a specific theme in the conception of his works on a macro-perspective level. Since the notebook entries also refer to artworks and artistic concepts, this method allows corresponding text passages to be searched for, in addition to the graphics. Only through the conflation of textual entries, graphical

Figure 8: Corpus-spanning search for shared concepts.

representations, external links and work manifestations, can a comprehensive analysis of the work creation process become possible.

6 Conclusion

Contemporary art, especially from conceptual artists, is usually not self-explanatory, nor easy to understand. From the 1960s onwards, artists dissociated themselves from the very notion of the artwork and the art market and progressed beyond the boundaries of conventional art production. They started to work in an interdisciplinary fashion and were no longer exclusively rooted in one single medium or genre, instead experimenting with different media such as photography, video, music, audio, language, and fine arts. The artist's actual opus is the concept *behind* a manifestation, a state in an ongoing creative process. The artistic concept becomes a reasonable art form itself, whereas the actual manifestation fades into the background. While little tangible evidence beyond memories from witnesses and documentation in catalogues and books remains of Skerbisch's early media installations, the records in the notebooks give testimony to the artist's intensive engagement with the conceptualisation of these installations and their leading topics.

A closer investigation of textual and graphical versions leads us not only to the core of an artwork, but more significantly to what determines art itself, more specifically:

the idea. Initially described by Platon and conceptually shaped in Antiquity, the questioning of art itself is central to conceptual art (Beyer 189).

Considering methods from other disciplines shows that versions exist in any kind of creative process. It is, however, necessary to differentiate versions, variants, and variations, and to identify the development stages and the different external influences in order to reveal components discarded during the design process, and finally the idea at the core of an artwork.

In the case of an artist, the final version is usually performed in another medium. Accordingly, a genetic approach needs to consider more than text to uncover the unfolding of an idea. As has been shown in this contribution, an extended genetic approach and the application of semantic technologies is needed to identify versions of the same conceptual roots. This facilitates the revealing of a network of initial ideas, temporal manifestations and continuous concepts, to help reconstruct the artist's creation process over time.

Looking at the notebook entries, it becomes obvious that the tent and the box were central shapes in Skerbisch's early media artworks, which were reused, reorganised and reconceptualised by the artist several times. The tent first occurred in the photograph *Patmos* in 1972. The tent and box shapes were then used separately by Skerbisch in his first exhibition at the poolerie in 1975 and subsequently reused and further expanded within the notebooks before he combined them in the installation *Erde (Our cubehouse still rocks)* in 1976. Furthermore, Skerbisch used the box in his first exhibition at the *poolerie* in 1975 and he combined it with the tent in *Erde*.

Coming back to the original differentiation of variant and variation, this means that we have several versions of tent constructions and several versions of box construction showing substantive and accidental changes, whereas the transition from one work of art to another work of art can be considered as a variation of a shared concept. The artworks in figure 9 are related by concept and content, but whereas the overall concept remains clear and consistent, single components vary to a large degree (Scholger, "Tracing the association processes").

Investigating the different versions of graphical representations of artistic ideas in notebooks—and especially the changes and alterations in material, positions, components, and proportions—supports us in solving the bigger puzzle of reconstructing an artist's creation processes. Such investigations should therefore be much more prominent in the genre of digital scholarly editing of artists' sources. To achieve that goal, however, it is paramount and required to comprehend and digitally represent graphics with the same depth and complexity as text.

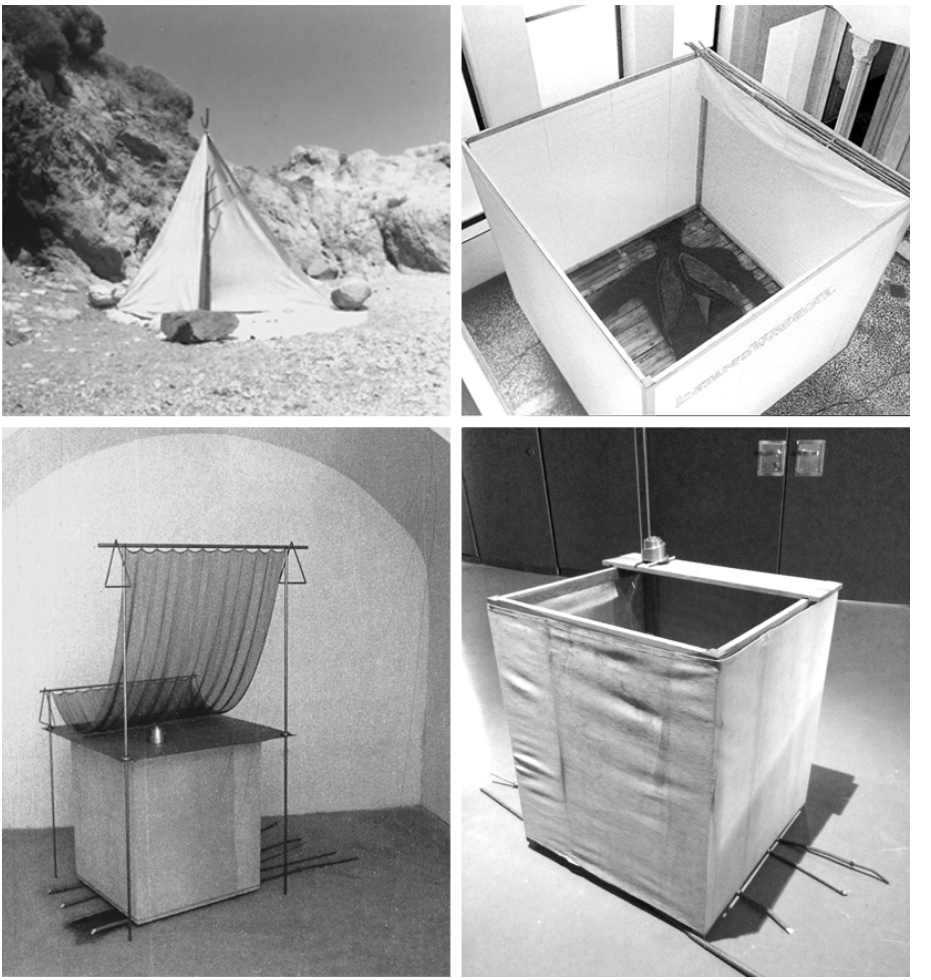


Figure 9: Patmos (1972), Erde. Our cubehouse still rocks (1976), Kasten 1975; photographs by Hartmut Skerbisch. Kasten (reconstruction) 2015.

Bibliography

Appel, Bernhard. R., and Joachim Veit. *Variante: Beethovens Werkstatt. Genetische Textkritik und Digitale Musikedition.* 2015, beethovens-werkstatt.de/glossary/variante-2-0-0. Accessed 8 Nov. 2018.

Beyer, Andreas. "Idea." Pfisterer, *Metzlers Lexikon Kunstwissenschaft*, J.B. Metzler, 2011, pp. 189–92.

Bowman, Frank Paul. "Genetic Criticism." *Poetics Today*, vol. 11, no. 3, 1990, pp. 627–46. doi:10.2307/1772829.

Burnard, Lou, et al. *An Encoding Model for Genetic Editions.* 2008–2013, www.tei-c.org/Vault/TC/tcw19.html. Accessed 5 Nov. 2019.

Cox, Susanne, et al. "Beethovens Werkstatt: Genetische Textkritik und Digitale Musikedition – Projektvorstellung." *Forum Musikbibliothek*, vol. 36, no. 2, 2015, journals.qucosa.de/e-journals/fmb/article/view/320/273. Accessed 24 Jun. 2019.

De Biasi, Pierre-Marc. "Horizons for genetic studies." *Word & Image*, vol. 13, no. 2, 1997, pp. 124–34. doi:10.1080/02666286.1997.10434277.

———. "Pour une approche génétique de l'architecture." *genesis*, no. 14, 2000, pp. 13–65.

Eggelhöfer, Fabienne, and Marianne Keller Tschirren, editors. *Paul Klee – Bildnerische Form- und Gestaltungslehre.* 2012, www.kleegestaltungslehre.zpk.org. Accessed 8 Nov. 2018.

Fenz, Werner. *Hartmut Skerbisch: Werkauswahl 1969–1994.* Neue Galerie, 1994.

Ferrer, Daniel. "Variant and Variation: Towards a Freudo-Bathmologico-Bakhtino-Goodmanian Genetic Model?" *Genetic criticism and the creative process: Essays from music, literature, and theater*, edited by William Kinderman and Joseph E. Jones, University of Rochester Press, 2009, pp. 35–50.

———. "Genetic Criticism with Textual Criticism: From Variant to Variation." *Variants*, 12–13, 2016, pp. 57–64. doi:10.4000/variants.284.

Fleming, Alexander. *Fleming Papers.* 1917–1918, British Library, Add MS 56148. www.bl.uk/collection-items/alexander-flemings-notebook-june-1917-1918. Accessed 24 Jun. 2019.

Flynt, Henry. "Essay: Concept Art." *An anthology: Of chance operations*, edited by LaMonte Young and Jackson Mac Low, 1963, www.henryflynt.org/aesthetics/conart.html. Accessed 24 Jun. 2019.

Greg, Walter W. "The Rationale of the Copy-Text." *Studies in Bibliography*, vol. 3, 1950, pp. 19–36. www.jstor.org/stable/40381874. Accessed 24 Jun. 2019.

Grésillon, Almuth. *Literarische Handschriften: Einführung in die "critique génétique."* Lang, 1999. Arbeiten zur Editionswissenschaft 4.

Hartmann, P. W. "Fassung." *Das grosse Kunstlexikon von P. W. Hartmann.* www.be-yars.com/kunstlexikon/lexikon_2787.html. Accessed 24 Jun. 2019.

Holler-Schuster, Günther. "Hartmut Skerbisch's work between media art and land art." *Hartmut Skerbisch*, Verein der Freunde von Hartmut Skerbisch, pp. 143–71.

Jansen, Leo, et al. *Vincent van Gogh. The Letters.* 2012, vangoghletters.org. Accessed 8 Nov. 2018.

Kosuth, Joseph. "Art after philosophy I." *Studio International*, vol. 178, no. 915, 1969, pp. 134–37.

———. "Art after philosophy II." *Studio International*, vol. 178, no. 916, 1969, pp. 160–61.

———. "Art after philosophy III." vol. 178, no. 917, 1969, pp. 212–13.
Künstlerhaus Wien. *K45-Kunst nach 1945: Ausst.–Kat. [Internationale Kunstmesse, Künstlerhaus Wien, 17. bis 22. Februar 1977]*, Gesellschaft Bildender Künstler Österreichs, 1978.
LeWitt, Sol. "Paragraphs on Conceptual Art." *Artforum*, 1967.
Leymarie, Jean, et al. *Die Zeichnung: Entwicklungen, Stilformen, Funktion.* Skira-Klett-Cotta, 1980.
Maugham, William Somerset. *A writer's notebook.* William Heinemann, 1949.
Music Encoding Initiative, editor. *Music Encoding Initiative (MEI).* 2019, music-encoding.org. Accessed 24 Jun. 2019.
Myssok, Johannes. "Bozzetto, Entwurfsmodell, Skizze, Vorstudie." Pfisterer, *Metzlers Lexikon Kunstwissenschaft*, J.B. Metzler, 2011, pp. 75–78.
Pierazzo, Elena. "Digital Genetic Editions: The Encoding of Time in Manuscript Transcription." *Text Editing, Print and the Digital World*, edited by Marilyn Deegan and Kathryn Sutherland, Ashgate, 2009, pp. 169–86.
Plachta, Bodo. *Editionswissenschaft: Eine Einführung in Methode und Praxis der Edition neuerer Texte.* 3., erg. und aktualisierte Aufl., Reclam, 2013. Reclams Universal-Bibliothek 17603.
Radecke, Gabriele. "Notizbuch-Editionen: Zum philologischen Konzept der Genetisch-kritischen und kommentierten Hybrid-Edition von Theodor Fontanes Notizbüchern." *Internationales Jahrbuch für Editionswissenschaft*, edited by Rüdiger Nutt-Kofoth and Bodo Plachta, vol. 27, De Gruyter, 2013.
———. *Theodor Fontane: Notizbücher. Digitale genetisch-kritische und kommentierte Edition.* 2015–2017, fontane-nb.dariah.eu/. Accessed 8 Nov. 2018.
Ristić, Predrag. *Geometry in Lepenski Vir*, 1970.
Rosen, Philipp von. "Fälschung und Original." Pfisterer, *Metzlers Lexikon Kunstwissenschaft*, J.B. Metzler, 2011, pp. 120–23.
Scheibe, Siegfried. "On the Editorial Problem of the Text." *Contemporary German editorial theory*, edited by Hans Walter Gabler, Univ. of Michigan Press, 1995, pp. 193–208. Editorial theory and literary criticism.
Scholger, Martina. "Tracing the association processes of the Artist – The artwork as a commentary." *Hartmut Skerbisch*, Verein der Freunde von Hartmut Skerbisch, 2015, pp. 305–19.
———. *Assoziationsprozessen auf der Spur: Digitale Edition der Notizbücher von Hartmut Skerbisch.* PhD diss., University of Graz, 2018.
———, editor. *Die Notizbücher von Hartmut Skerbisch. 1969-2008. Eine digitale Edition.* 2018, gams.uni-graz.at/skerbisch. Accessed 25 May 2018.
———. "Taking Note: Challenges of Dealing with Graphical Content in TEI." *Journal of the Text Encoding Initiative*, Issue 12, journals.openedition.org/jtei/.
———. "Thesaurus for Graphics." 2018, gams.uni-graz.at/o:sker.graphics. Accessed 25 May 2018.
Shillingsburg, Peter L. *Scholarly Editing in the Computer Age: Theory and practice.* 3rd ed., University of Michigan Press, 1996. Editorial theory and literary criticism.
Sykora, Katharina. *Das Phänomen des Seriellen in der Kunst.* 1983.
Systap. *Blazegraph.* 2016, www.blazegraph.com/. Accessed 8 Nov. 2018.
Stein, Gertrude. "Sacred Emily". *Selected Writings of Gertrude Stein.* Vintage Books, 1990
Tanselle, G. Thomas. "Greg's Theory of Copy-Text and the Editing of American Literature."

Studies in Bibliography, no. 28, 1975, pp. 167–229. www.jstor.org/stable/40371615.

TEI Consortium, editor. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.4.0.* 2018, tei-c.org/Vault/P5/3.4.0/doc/tei-p5-doc/en/html. Accessed 8 Nov. 2018.

The Getty Research Institute. *Art & Architecture Thesaurus Online.* 2015, www.getty.edu/research/tools/vocabularies/aat. Accessed 8 Nov. 2018.

The Interedition Development Group. *CollateX – Software for Collating Textual Sources.* 2010–2018, collatex.net/. Accessed 8 Nov. 2018.

Turner, Jane. *The Dictionary of Art: Diploma work to Egypt.* Grove, 1996 9.

———. *The Dictionary of Art: Medallion to Montalbani.* Grove, 1996 21.

———. *The Dictionary of Art: Savoy to Soderini.* Grove, 1996 28.

Verein der Freunde von Hartmut Skerbisch. *Hartmut Skerbisch: Leben und Werk, life and work: Gegenwart als Gegenwart, present as present.* Verlag für Moderne Kunst, 2015.

Versioning Cultural Objects through the Text-Encoding of Folk Songs

Richard Breen

Abstract

This paper will present and discuss experiences studying different versions of folk songs as cultural objects, and will investigate how using specific Digital Humanities tools may assist the versioning of intangible oral tradition. This was primarily achieved using The Versioning Machine, a framework and an interface for displaying multiple versions of text and audio encoded according to the Text Encoding Initiative (TEI) Guidelines. Through encoding a set number of songs in The Versioning Machine and displaying the results online, new questions and conclusions could be made to version cultural material with an emphasis on trying to trace the evolution of cultural ideas through subsequent iterations of ideas. Using examples from the project *Documenting Transmission: The Rake Cycle*[1], this paper will examine the effectiveness of using a specific existing versioning tool to model and map the differences between versions of folk songs and examine the intangible nature of performance and oral tradition. How do these digital versions change or reinforce our perception of a song cycle and transmission processes in general? This paper will give a broad overview of the Documenting Transmission project and some of the musicological and technical considerations that were made over the course of the project.

1 Introduction

The malleability with which the term *version* is used across disciplines within the humanities is certainly valid cause for careful evaluation to assert some form of working definition across them. The advent, and continuing growth, of inter-disciplinary scholarship forms new perspectives, but also informs the treatment of humanities data. Given the vast amount of different media through which humanities data may now present itself, a broad treatment of what the term *version*, and the subsequent process of *versioning*, is should be paramount. Ontologically, the attempt to digitally classify a group of related cultural artefacts as versions of one another is hindered

[1] A link to the project may be found in the bibliography. N.B: for best functionality, please view in Mozilla Firefox.

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 57–75.

firstly by the unquantifiable nature of much humanities data. With cultural material, particularly that of oral tradition, authority is often lost through ill-defined chronology. Attempting to digitise this intangible cultural heritage is a vast undertaking, particularly with vast amounts of information with often-loose structure based on oral tradition. Despite the ontological woes of dealing with humanities data and its intrinsic differences to scientific data, the treatment of cultural material in a digital environment at least allows us to re-approach older concepts, asking old questions and finding new answers. In an ethnomusicological context, we may begin to digitise and map out tradition in new ways. Approaching versioning as a process is an understanding that the contextualisation of a cultural artefact is essential to the perception of the subsequent versions of that object. The modelling of these processes is therefore becoming an increasingly relevant issue among Digital Humanities scholars. The digital offers myriad new ways in which one may perceive artefacts as versions of an original artefact, and digital scholarly editions help embellish the narrative of an object's history and help form new perspectives for that history's treatment.

This paper will address versioning cultural objects through the case study of the *Documenting Transmission: The Rake Cycle* project, which text-audio encodes folk songs from the well-known "Rake Cycle" and visualises the transmission processes that occur between the songs in the *Versioning Machine* (Schreibman et al.). The term "Rake Cycle" refers to a nineteenth century English folk ballad entitled "The Unfortunate Rake," which through transmission processes has evolved into many variants in several sub-genres of music. While there are variants in languages other than English, this project text-encoded the lyrics and audio of thirteen English-language versions from Kenneth Goldstein's "The Unfortunate Rake" album, chosen for their motivic, musical, and geographical similarities and differences. The intention of this project is to highlight and display motivic and tropic information across the texts and music as opposed to simply the music itself. Previous and ongoing digital versioning projects, such as *The Thomas MacGreevey Archive* and *In Transition: Selected Works by the Baroness Elsa Von Freytag Loringhoven* (Clement) that have used text-encoding to mark-up versions of poems have dealt with the relationship between witnesses to one overarching textual idea. Tanya Clement, when referencing the *In Transition* project, describes the relationships between draft versions of poems as *Textual Performance Theory*, wherein the view is taken that the relationship and dialogue between each version of each poem is as important as the final product (Clement, "In Transition"). There is no definitive version of each poem, it is the differences between them that becomes interesting, whether it is the semantic differences between stanzas or instances of words, or the visual aspect of the words on the page themselves (Clement, "Knowledge Representation" 3). The *In Transition* project presents twelve unpublished poems written by Freytag-Loringhoven and marks them up according to the Text Encoding Initiative's P5 guidelines. The poems are then represented and edited critically

in the *Versioning Machine*, a customisable framework used to display text encoded according to the TEI guidelines (TEI Consortium), facilitating the close comparison of several texts, or *witnesses*, with a diplomatic or authoritative text (Schreibman et al., *Versioning Machine 5.0*, "Documentation"). Each poem contains several different *sketches* of that particular poem at different draft stages, the point of which is tracing the genesis of each poem from the earliest sketch to the completed poem. This assumes a process wherein the relationship and dialogue between each version of each poem is as important as the final product. Susan Schreibman, creator of the *Versioning Machine*, states that it "[e]nables […] a theory of social-text editing wherein no version of the work carries more authority than another: each version of the work being a witness to a textual moment" (Schreibman, "Re-Envisioning" 93). The texts and sketches that are formative to a completed piece of work are therefore seen as a process that is as important as the result. This provides a complete textual history rather than establishing a definitive text. In marking-up literary work, a new digital version of said work is also created.

In classical music study, similar research has been made in *sketch studies*, wherein a composer's draft manuscripts are observed relative to a completed work to form a genetic story of that work (Kerman 174). Some digital projects have begun the migration of this theory to the digital, marking up scores according to the Music Encoding Initiative (MEI) Guidelines. The *Beethoven Werkstatt* (Appel et al.) employs an MEI-based digital mark-up of the manuscripts of Ludwig van Beethoven. Similarly, the *Online Chopin Variorum Edition* (OCVE) marks up the many published versions of Frédéric Chopin's scores to compare publication histories and the minute publication differences between the composer's scores. Perry Roland explains that the true potential of the MEI and its implementation in the OCVE project is to "encode multiple versions of a musical work and generate multiple outputs" (Roland 8). On a functional level, this is a very similar principle to that of the aforementioned *In Transition* project; it is the addition and subtraction of material in Chopin's differing printed editions that generates an interest in documenting these different editions of that composer's work. The result is that it becomes much harder to call any one of Chopin's printed scores as a *definitive* work.

Musically speaking, the *Documenting Transmission* project was concerned with the digitisation and versioning of oral material, primarily through the transcription and encoding of lyrics and audio tracks from an album of folk songs that display what is known as the *Folk Process*. As folk songs travel geographically or are shared culturally, they may adopt completely new musical characteristics while maintaining a core lyrical or thematic story. Norm Cohen states that:

> All folk ballads are distinguished by a tendency of singers or composers to reuse stock phrases and even entire stanzas from older ballads […] borrowing

> phrases and stanzas from earlier works is no more plagiarism than it would be for a *literary* poet to hunt for words in a thesaurus…

One such issue with this project was that none of the resulting variations of the ballads of the "Rake Cycle" can be considered as the definitive version due to most of the changes in either music or lyrics being subject to the process of oral transmission. The focus is not on one author or composer, but rather many different musicians re-interpreting and building upon a source material to create their own version. Where *Documenting Transmission* differs is that where MEI-centred projects are concerned with the marking up of set musical notation, this project is concerned with the comparison of textual differences in lyrics, with the music serving more as an aural guide to the lyrics and cultural motifs. This is more demonstrative of the transmission process than that a formal analysis of the music itself, in that there is in most cases no formal or authoritative notation of the music in question. As such, the TEI Guidelines and Text Critical Apparatus tag set is used instead of the MEI Guidelines to encode the lyrical texts, and the *Versioning Machine* is used to display the results.

There are key points to consider when versioning such texts, namely that:

1. Folk songs are performances and are generally not transcribed lyrically or musically as in the Western Art Music tradition.
2. Attempting to map oral tradition presents a huge problem of authority between versions often due to a lack of publication or performance history.

While there are countless digital repositories dedicated to the digitisation and preservation of folk music and material, such as the *Vaughan Williams Memorial Library* (VWML) or the *Comhaltas Traditional Music Archive*, there is often a tendency to amass material and less of a focus on the digital modelling and study of the relationships between songs in the vein of a digital scholarly edition. This project was therefore intended to marry the methodologies of ethnomusicology and previous text-critical versioning projects such as those listed above. While it would be possible to adapt *The Versioning Machine* to display MEI, the sheer size of musical variation between many of the versions used in this study means that attention to minute musical change is eschewed in favour of demonstrating broader motivic changes in lyrical tradition.

The most recent release of VM contains a new text-audio linking feature, which had originally been developed by members of the Modernist Versions Project and is now a standard component of VM 5.0.[2] This feature facilitates parallel reading of a version of a text and at the same time listening to an audio version. In observing how other projects utilised the *Versioning Machine* to visually represent the versioning process, *Documenting Transmission* utilises and extends some of these features of VM

[2] See v-machine.org/documentation/#enc_audio.

5.0 to allow the comparison of different motivic, lyrical, and musical features of folk songs.

2 The folk process

The study of folksongs as an example of the tracing of cultural lineage between versions is a long-studied area of musicology and ethnomusicology, for at least over one hundred years. In 1907, the seminal English musicologist Cecil Sharp noted that folk music is communal in two forms, in that of its creation and in its representation of the thoughts of the community. In *English Folk Song: Some Conclusions*, he quotes F. M. Boehme, stating "[f]irst of all one man sings a song, and then others sing it after him, *changing what they do not like*" (Sharp 10). Perhaps the most famous of Sharp's observations in this book is his identification of three factors that govern the transmission of folk songs musically and the forms these songs took depending on their context. These were:

1. Continuity: the linking of the past and present.
2. Variation: that the changes that occur rely solely on the creative tendencies or impulses of an individual or group.
3. Selection: that the community in question chooses which music it plays, which in turn decides the form(s) the music takes as it is survives. (Sharp 17–31)

Sharp's observations are an effort to identify how internal and external factors effect or determine the direction of the process of oral transmission, such as war, politics, and societal change. Musicologists throughout the mid to late 20th century, while being undecided as to how exactly to define folk music, agreed that Sharp's three characteristics were intrinsic to the creation of folk material (Cowdery 808). Charles Seeger is credited with coining the term *folk process* to define the "…process by which cultural artefacts [sic] are changed, whether minutely or in significant amounts, to form new cultural products" (Washbourne 457). A group of folk songs with similar tunes or similar lyrics may therefore be identified as being a part of the same song family, wherein each performer identifies their own version of a song.[3]

The *Documenting Transmission* project centred mainly on trying to digitally map the folk process in action across selected songs from a song cycle to show the transmission and transformation of literary and artistic material from person to person in both an

[3] David Atkinson, speaking of the genetic links between folk songs, states that "a particular song as taken down from a particular contributor is most usually said to constitute that person's version of that particular song type, wherein *type* means the constant elements that "unify […] what is recognizable as the 'same' song" (4). It should be stated then that this definition of version is based around the folkloristic interpretation of a version. Since this project is concerned with representing this definition of a version, as opposed to representing textual witnesses, this definition of version is used when describing individual songs.

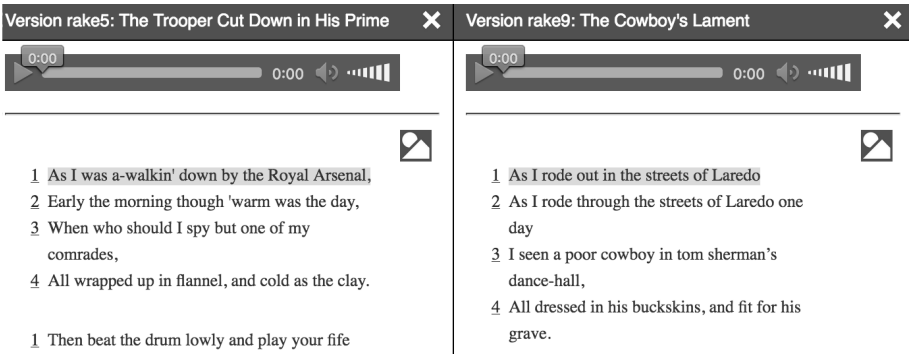


Figure 1: How songs, or Versions, are presented in the *Versioning Machine*. A user may press play on any line of text and hear the audio of that corresponds to that line of text. Similar lyrical ideas are highlighted across all versions corresponding to whichever track is selected.

oral and literary context. Essentially this means that as the stories, art, and musical traditions of a community are passed down from generation to generation, they are subject to organic changes in form, context, narrative and performance. *Documenting Transmission* is an investigation into displaying the interoperability of the narrative, musical, and visual functions that song lyrics serve. In some cases, certain elements may be completely changed or left out completely between versions or the music may be entirely different. In a digital context, this raised the question of whether existing DH tools could be used to reflect both musical and literary change in a way that extends our consideration of these processes. The *Versioning Machine* framework allows a user to deconstruct and observe similarities and differences between versions concurrently, in such a way where a user may be aware of, or disregard, their historical context at will. They may press play on any one of the versions and actively see where lyrical similarities or differences occur in real-time across every version (see figure 1).

The collection of songs used for this project comes from a cycle of which the earliest surviving version is a fragmented verse from a 1790 broadside pamphlet, entitled "My Jewel, My Joy" (Lodewick 98). The subsequent full, or more substantial, versions in the cycle have been the subject of analysis for quite some time, with major ethnomusicological criticism coming in the mid-twentieth century by Kenneth Goldstein and Kenneth Lodewick. The most popular of the early versions of the song dates from 1840, most commonly titled "The Unfortunate Rake" or "The Unfortunate Lad" (Harwood 26; Roud Broadside Index B130326). The versions that were encoded for the project are taken from Goldstein's educational compilation album *The Unfortunate Rake* released in 1960 by Smithsonian Folkways, its intention being a "study in the evolution of a ballad". This album's liner notes provide much background information

about each version and provide transcriptions of the lyrics, as well as contextual information for the performances. Encoding these songs as versions therefore required the encoding of both the lyrics (content) and the audio (performance) in the form of MP3 files.

The traditional story of this homiletic ballad recounts a protagonist who happens upon a former comrade dying of venereal disease at the side of the road. Subsequent verses detail how the invalid came to be "disordered" by a camp follower, assumedly a prostitute (Goldstein, "Liner Notes" 2). Depending on where the ballad travels to, the role of the soldier changes; it travels to sea and becomes about a sailor, or to the Americas to become a cowboy or miner. Other functions of the lyrics such as places, people, or events also change depending on that version's environment or performer. This lyrical story and its countless reinterpretations of the lyrics form certain tropes and possibly hundreds of various incarnations and parodies across the world over a period of at least two hundred years. This identification of tropes and motivic functions in ethnomusicological terms is rooted in anthropological and folkloristic study. In his seminal *Morphology of the Folk Tale,* the folklorist Vladimir Propp made efforts to break down the basic plot components of Russian folk tales to reduce them to their most basic functional state. He observed that fairy tales from Russia were made up of thousands of literary functions that interacted with one another in countless combinations throughout the canon of Russian folklore. This fluidity of actions as opposed to unchanging material applies directly to the folk process and oral traditions. Propp argues that actions are more important than the dramatis personae, this too can be said of the Rake Cycle (7). The invalid in any of the folk songs may be a soldier or a sailor, his gender may be reversed, or the setting of the story may completely change. What is important is that there are certain constants and variables that may be found in every version. These include a military funeral, an invalid, or a warning to the listener. The *Versioning Machine* is therefore well suited to show these changes, as many small minute links can be location-referenced in context to one another.

3 The *Versioning Machine* and the text-encoding process

The *Versioning Machine's* primary function is to display text encoding to facilitate close comparison of several texts, or witnesses, with a diplomatic or authoritative text (Schreibman et al., *Versioning Machine 5.0*, "Documentation"). It does this by taking a TEI document and transforming it into a HTML page that can be opened in a browser. Some features include the ability to add annotations and notes to encoded text, as well as providing an image viewer that can pan and zoom. These features were utilised in the *Documenting Transmission* project, with further aspects being implemented such as a colour coded motif reference index. This basic visualisation was used to identify

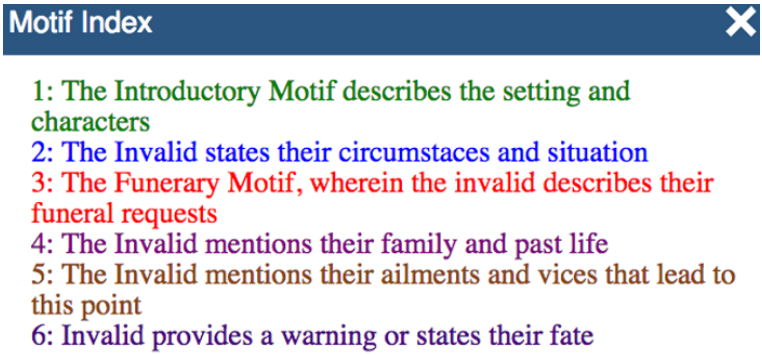


Figure 2: The Motif Index indicates which motifs are assigned which colour.

different motifs and thematic ideas across versions. This was achieved by customising the *Versioning Machine* by adding a simple plugin that places two buttons on the *Versioning Machine's* interface. These toggle on and off a visualisation of the motifs across the song cycle by highlighting certain stanzas or lines in a pre-determined colour (see figures 2 and 3).

The Critical Apparatus tag set is designed to provide editors with a structured method of recording differences or variations between multiple witnesses of the same text.[4] Using this tag set allows an editor to encode multiple versions of a text in a single document. A temporal alignment of text and audio is created using location based referencing, wherein the text of a transcription is encoded in the TEI critical apparatus structure. The audio files themselves are declared in the TEI `<witDetails>` element. Via its `@target`, an audio file can be linked to a particular text version.

```
<witDetail wit="#rake1" target="#rake1">
  <media mimeType="audio/mp3" url="audio/unfortunate_rake.mp3"/>
</witDetail>
```

The Versioning Machine's XSLT stylesheet detects if an audio file was aligned with a text and includes an HTML audio element at the transformation stage into the VM interface. The visual alignment between different parts of the audio track and the transcribed text happens via a jQuery plugin. By clicking on a section of a text, the section is highlighted and the corresponding audio is played. Additionally, TEI `<timeline>` and `<when>` elements are embedded to provide a set of ordered points in time. Every `<when>` element represents a piece of text and is linked to a fragment of an audio track. In that way a temporal alignment of text and audio is achieved.

[4] See v-machine.org/documentation/#method2.

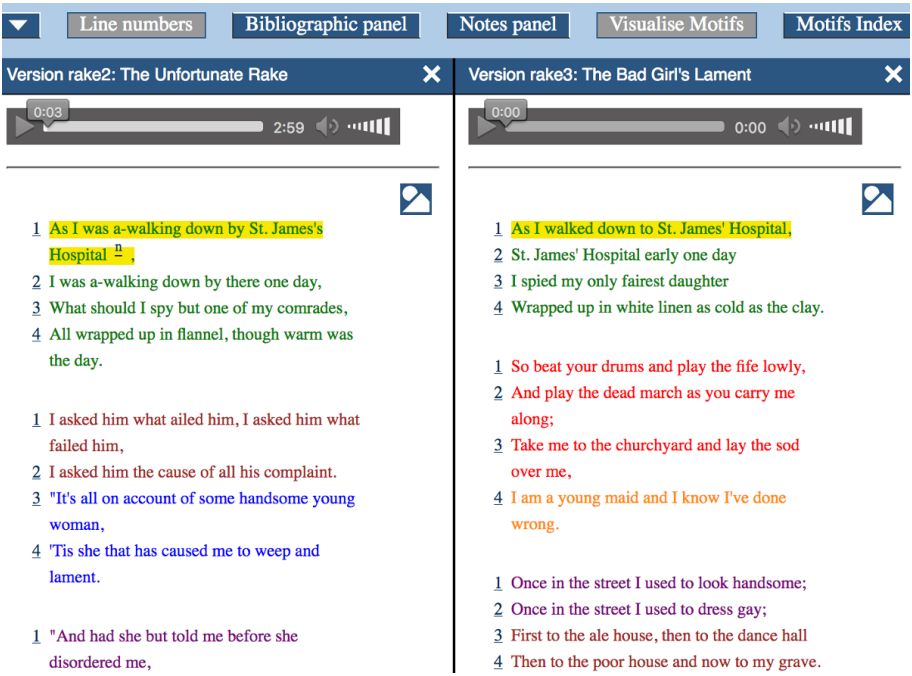


Figure 3: Motifs highlighted across two versions.

```
<l n="2">
  <app loc="locA1">
    <rdg wit="#rake3">
    <timeline unit="s">
      <when since="#tl_rake2_2" interval="4" xml:id="tl_rake2_3"/>
    </timeline> St. James' Hospital early one day
   </rdg>
 </app>
</l>
```

In the code example above, *locA1* refers to a specific point in time that will correspond with another set of lyrics identified as *locA1* in another version. This is essentially how two pieces of text are identified and highlighted at the same time. By clicking on a section of a text, the section is highlighted and the corresponding audio is played. A user may press play on any of the encoded versions and, while audio is played on that version, corresponding or similar lines of text are highlighted across all versions in real time until a song ends. Alternatively, a user may press on any line from any version to hear that version instantly (see figure 4). Allowing a user to break each song up and listen to any line of any version at a given time in any combination,

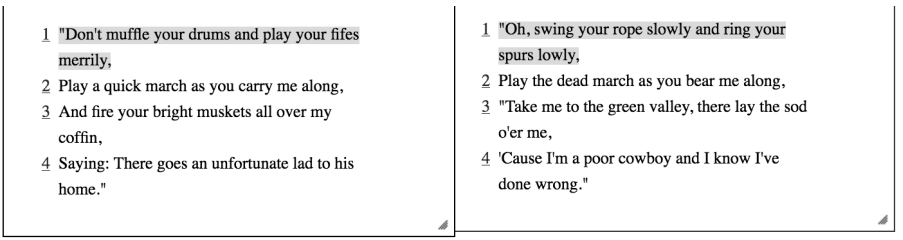


Figure 4: The Versioning Machine allows a user to see where similar motifs occur across different versions.

independent of the context. In this way, new spontaneous *songs* may be formed, albeit with a sense of chance. It can be said a user is actively *remixing* the song cycle, creating new cultural objects in the tradition of the song cycle, brief though they may be.[5]

The lyrics for each song were first analysed alongside one another for any immediate similarities by hand. The first level of analysis was to observe each set of lyrics and compare them to one another to find similarities or differences. These were informed largely by a number of texts that have analysed the song cycle, predominantly Goldstein's album liner notes and Lodewick's "The Unfortunate Rake and His Descendants." In the case of this song cycle, according to Goldstein ("Liner Notes" 1) and Lodewick (99), the earliest surviving documentation of the ballad is a single surviving verse titled "My Jewel, My Joy." This was encoded as the first version and serves as a starting point to trace motivic development across subsequent version. As this does not have an audio track on the album, there is no audio for this in the project. The first version encoded with audio is Track 1, "The Unfortunate Rake." Goldstein asserts that this version is "sufficiently close enough to the original ballad to warrant its use as a starting point for an examination of the whole family of related parodies and recensions" (2). As such, this project used "My Jewel, My Joy" and "The Unfortunate Rake" as a starting point from which changes can be observed across multiple versions of the ballad.

4 Observations

In figure 4, the left column displays a version of the ballad in its oldest recorded form "The Unfortunate Rake," whereas the version on the right, "Streets of Laredo," is recorded nearly a century later. The lyrics of the version on the left are more

[5] The TEI document for this project is available on GitHub at github.com/rudgebreen/DocumentingTransmission--The-Rake-Cycle/blob/44211aa243c7973c956aab710b29580ebc9123c1/RakeCycle.

antiquated, depicting a military setting. The lyrics of the ballad on the right illustrate the relocation and recycling of the funerary motif to suit a cowboy context. Clearly some lyrical constructs have been carried over from the older song on the left and have been transmitted into that of the right, signalling that the action that is carried over in the cycle is more important than the song making historical or contextual sense.

In allowing a user the ability to click anywhere on any given version at a time, the *Versioning Machine* demonstrates a *deformance* of these encoded songs (Schreibman 99). Essentially, the performative meaning of a text is broken by way of taking the text out of its syntactical context. However, this project not only breaks the syntactical context of the lyrics visually. The entire tonality of the reading is changed through the changing of the music in tandem with the changing of the lyrics. This does not necessarily mean that the aural and visual readings are always broken together. There may be a line in one version that is completely the same syntactically as in another version, but the corresponding audio of each version is totally different. For example; both Version 3, "The Bad Girl's Lament," and Version 12, "The Wild Lumberjack," contain the identical lyric: "And play the dead march as you carry me along." However, the audio for Version 3 contains a voice accompanied by a guitar playing in a minor key with a definite pulse, whereas Version 12 contains one unaccompanied voice performing in a major key.

In terms of basic musical analysis, these are two completely different pieces of music regarding melody, tonality, and timbre. Observing just the lyrics by themselves creates no narrative context, it is the playing of the music that gives it some sort of context to the listener subconsciously. This is an immediate reaction as opposed to one that is informed by any previous textual information, and can be expressed without any need of the rest of the text. The melody that a singer chooses or the tonal quality of their voice can inform much to a listener about that line of text. The singer may be portraying conviction, sorrow, happiness, et cetera. Through this one lyric and using only the audio as the distinguishing feature between the two a further emphasis is placed on the performative aspect of the text and it adds to the sense of repurposing of the lyrics. If the folk process is the reshaping of cultural material into new cultural artefacts, then the performative nature of this project is as important to the result as the lyrics themselves. The encoding of audio alongside the text is essential in giving the user context both consciously and subconsciously.

The *Versioning Machine* allows the user to see the strong correlation between content (lyrics) and form (performance) in each of these songs. This in turn highlights the importance of influence between these songs and their performers. This has a huge amount to do with cultural transmission. As the ballad travels to new places, slang is re-used and misinterpreted to the point where new words and readings are created. One such example is "Bright Summer Morning," recorded in the Virgin Islands. In

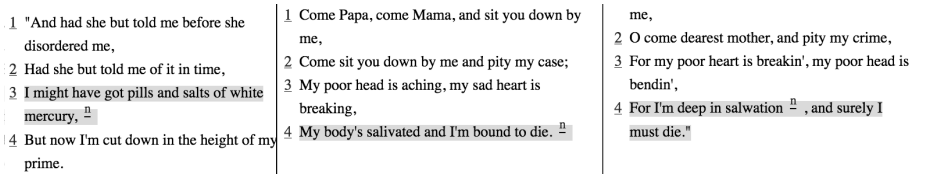


Figure 5: Each Version alludes to venereal disease in their lyrics, although the genders are reversed in the middle and far right versions.

this version, the over-arching story changes perspectives of the protagonist to that of a prostitute who has been wronged by a man. In this recording, the performer describes that her body is in "deep in salwation." This is a corruption of the term "salivated," which Goldstein describes as mercury poisoning in the album's liner notes (4). Both versions appear to have come from "The Unfortunate Rake," wherein the Rake complains that had the prostitute disclosed that she had a venereal disease before "disordering" him, he could have "[...] got pills of salts and white mercury." Despite changing the perspective of the protagonist of the song, "One Morning in May" and "Bright Summer Morning" keep the source of the characters' "disordering," in doing so it re-shapes the lyrics to tell a new story and changes the listener or reader's perspective in the process. This is contextualised by encoding these notes in the TEI, which the *Versioning Machine's* interface displays (see figure 5).[6] Corruptions of lyrics in this way demonstrate this process of change, the performer is unwittingly creating a new cultural artefact that the user may observe happening in real time.

While influence and tradition clearly play a massive role in how each version of the folk-song and its successive descendants are presented, a question arises as to the individuality of a version or performance. At what point does a performer's version become recognised as its own song? Nearly every one of these versions of the story has a different melody or structure, and yet they are still related whether through a fragmented refrain or motif. Many of the similar verses or lyrical motifs may appear at many different stages of a performance. It is not fair to assume that every one of these performers knows the heritage of the ballad or where the story elements come from. To what extent then do these songs remain in the same tradition and when do they form new musical heritage?

One such example is the final version encoded in this project, "Gambler's Blues," which marks a shift toward the jazz and blues tradition that would eventually become the jazz standard "St. James Infirmary Blues." The lyrics to both versions contain only scant reference to the 18th century Irish ballad (Harwood; Lodewick). The

[6] Goldstein asserts that the song probably made its way to the Virgin Islands by way of British colonisers in the 19th century, when the islands were under English control ("Liner Notes" 4).

ballad has been re-purposed to the point where it breaks away into a new musical tradition at the genesis of a new genre. As Lodewick states, the historical lineage of this particular version is blurred to the point where no real conclusions of relation may be drawn outside of basic motivic evidence (Lodewick). While this may be true, the *Versioning Machine* facilitates real-time comparison of this version of the song against the others in such a way that a user directly sees and hears these differences. A user may observe evidence of the influence of oral tradition on performers. By transcribing these lyrics into the digital format, the encoding process is a continuation of the process undertaken by Goldstein; in creating the *Unfortunate Rake* album, he transcribed the performances themselves, including the slang used by the performers, which he then contextualised critically. In encoding this process, the project is therefore creating a new cultural object.

While the *Versioning Machine* serves the purpose of comparing each of the sets of lyrics and audio from each song, this only displays the folk process at a textual and musical level through the thematic differences the lyrics represent, and the musical differences assumed by each performer. This facilitates close observation of the characteristic differences between versions on a micro level. Visualising the data is done in part by providing a colour coding of the motifs across the lyrics in the *Versioning Machine.* It was also important to try and use a tool that would map these changes at a macro level. This was done by mapping each of the versions used in the project in *StoryMapJS* to show the geographical distance over which the song cycle accrues these cultural motifs and tropes that define it. *StoryMapJS* is a free online tool that allows a user to tell stories on the web that highlight the locations of a series of events (Knight Lab). Through mapping where in the world each version of the song is iterated, new conclusions about the motivic links observed in the *Versioning Machine* may be drawn based on where the ballad travels to. A large element of the folk process is typified by where music travels to and the changes that can occur as music is transmitted through different cultures. *StoryMapJS* was useful in representing this, as it gives some locational indication of how far the ballad travels over the course of the versions used in this project. It also facilitated the implementation of images of the lyrics from each of the ballads and contextual information to be inserted alongside these for the user. As most of the ballads are either biographically situated by Goldstein in his liner notes or are directly in the lyrics themselves, this map was intended to give a broad indication of the song cycle's migratory span.

This is not an exhaustive representation of every version of the songs from this cycle. It serves as more of a contextual tool to embellish the section of the project that uses the *Versioning Machine* (see figures 6 and 7). The user is presented with bibliographic detail about each version of the ballad but also is provided with the lyrics above this information (see figure 7). Through the story map the user is brought on a curated geographical journey of the history of the song cycle to supplement the

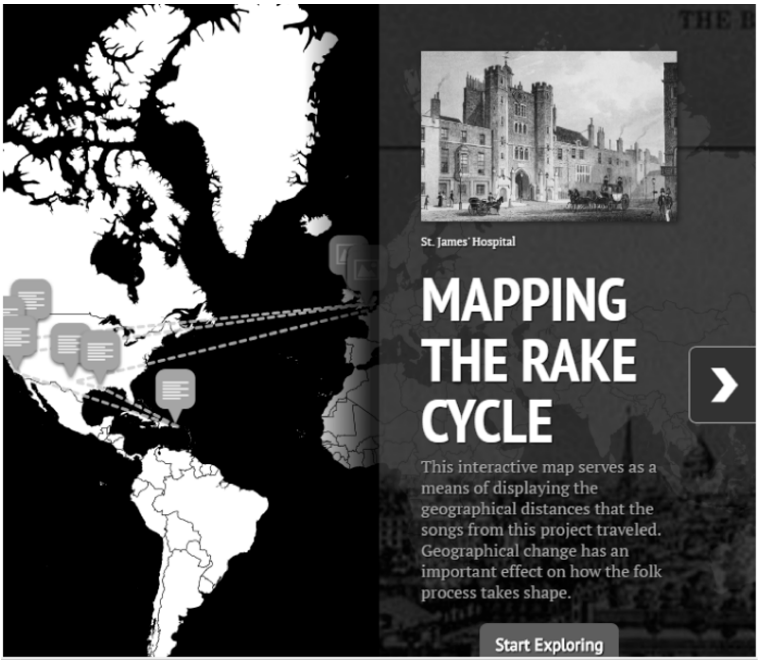


Figure 6: StoryMap JS acts as a curated guide through the migration of the song cycle.

material in the *Versioning Machine*. What is apparent from this visualisation is that after the song-cycle makes its way to the US, there is no exact traceable direction that the ballad takes, outside of thematic differences, which can also be inferred from studying the cycle in the *Versioning Machine*.Examples of this are Versions 8 and 10, "St. James' Hospital" and "The Streets of Loredo." Despite both variants being recorded in the same state in the US—Texas—they are vastly different in terms of content. What this visualisation indicates is that no matter where it is, these motifs are travelling to at a given time, they are transmitted by the performers based on their own intent. This is demonstrative of the melting pot of cultural differences that typify a huge country such as the US, and is interesting to realise this visually outside of the *Versioning Machine*.

5 Conclusions

This project can be viewed as a prototype for musicological study of folk songs in a digital space. The most beneficial aspect of this is the provision of a unique way

Figure 7: StoryMap JS allows key points to be mapped to give a broad idea of distances travelled, as well as the inclusion of biographical information.

of observing motivic changes and representations through using TEI encoding and visualisation software. The motifs that carry over from performer to performer are indicative of the pervasiveness of certain themes or ideas that typify this song cycle, and viewing this process on a broader level gives an insight into how these processes function as part of a larger commentary on transmission processes.

One of the major aspects this project could expand on and address is the sheer amount of musical variation between versions. While technically both the music and lyrics of each version are encoded, the software itself limits how actual musical differences between versions are represented. These complications make it hard to digitally model the intrinsic link between the lyrics and the music outside a purely aural indication that is up to the listener. Some visual indication of the musical differences would be extremely beneficial in representing this aspect of folk song. While this does not affect the overall research question of the project, it does provide an avenue for later research into the representation of musical differences between folksongs to demonstrate musical change. This text-encoding method therefore suits

the research question in that illustrates motivic and tropic change across both text and music. However, the adaptation of the *Versioning Machine* to display different encoding guidelines such as MEI would certainly be an interesting area in terms of the close-reading of sheet music. Authority between versions is also an issue: in some cases, with certain versions of these folk songs very little is known regarding their date of composition or from which regional tradition each variation of the ballad is coming from after a certain point in a song's emigration. Other music-centred projects, such as the *Online Chopin Variorum Edition*, wherein the cultural material represented takes the form of published music, is aided through historical publication as well as dating the primary manuscripts themselves. In this way, definitive chronology of the material may be asserted. Mapping the data in this project does not lead a user to definitive historical conclusions, but it does help clarify how influence and tradition affect transmission processes.

The utilisation of location-based referencing caused some issues throughout the encoding process, not because of the functionality itself but because of the encoding of both audio and text. Location-referenced encoding in the *Versioning Machine* was easily the most time-consuming aspect of the encoding process. The reason for this is that because this project requires each individual line of text from each separate version to align with a specific place in each of the audio tracks, each line of text then needed to be encoded according to a specific time in the audio file between the previous and following lines of text. This is not an issue in purely text-based projects, as you can display how witnesses change across different lines of the text without the need of linking text to a defined section of audio, multiple lyrics that were similar could have been encoded in the same apparatus `<app>` element (Schreibman et al., *Versioning Machine 5.0*, "Documentation"). Some of the elements or tags that make up the tag set contradicted the research being done syntactically. For example, each version was encoded with the `<witness>` element, although this research asserts that these versions cannot be labelled as *witnesses*. This was worked around by changing how the versions were displayed in the actual interface itself by editing the HTML file, which allowed the *Witness List* to be changed to *Version List*, and each *Witness* could be changed to *Version* (see figure 8). However, in the TEI document itself the versions are still labelled as witnesses. This is indicative that in this case, extending the TEI ODD in the future to better reflect the mark-up of these versions in the *Versioning Machine*.

There are certainly benefits to exploring transmission processes through observing the inheritance across folk traditions in digital spaces. The pedagogical potential of creating digital editions in this vein would certainly help to give musicological study more presence in the Digital Humanities. While this can certainly be refined and developed in the future, it is indicative that the TEI can be used to identify and represent the links between versions that are not formally published or part of larger

editions. This project demonstrates that motivic elements of folk songs can become popularised to the point of becoming tropes, and that these tropes become more and more representative of a song cycle as it ages and travels. What is reinforced is that this type of multi-media based tool has worth within the humanities for explaining how different aspects of intangible concepts such as the folk process can be both contextualised and represented in a digital space.

Mapping out this song cycle in the *Versioning Machine* gives a greater sense of visual and aural contextualisation regarding these recordings as cultural objects. This method of modelling and treatment of sources is not restricted to any one case study or song cycle. This research demonstrates how creating resources that contextualise and visualise transmission processes can display the shaping of a musical or literary tradition, particularly in ways that apply to enthusiasts in musicology or folklore studies. In identifying changes and similarities on a small scale through location-based encoding in the *Versioning Machine*, a close reading of the texts and performances is facilitated. The creation of a visualisation through colour coding provides an aspect of distant reading, wherein the user can trace these motifs as larger groupings of narrative. Presenting the larger bibliographic narrative and migratory patterns of the song cycle in *StoryMapJS* allows the user to spatially contextualise it. Through using these different technologies in tandem, transmission processes can begin to be mapped within the digital spectrum.

Bibliography

Appel, Bernhard R., et al. *Beethovens Werkstatt*. beethovens-werkstatt.de/. Accessed 12 Nov. 2017.

Atkinson, David. "A Critical Edition of Folk Songs and Plays: Imaginings and Constraints." Presentation at the *Book History Research Network Study Day on Electronic Texts*, Institute for Textual Scholarship and Electronic Editing, University of Birmingham, 2006, pp. 1–13. www.abdn.ac.uk/elphinstone/documents/Birmingham.pdf. Accessed 12 Nov. 2017.

Breen, Richard, editor. *Documenting Transmission: The Rake Cycle*, 2016. dhprojects.maynoothuniversity.ie/rbreen/Documenting%20Transmission:%20The%20Rake%20Cycle/samples/unfortunateRake.xml. Accessed 30 Mar. 2017 [Unfortunately the website is not working anymore. The source code is available on GitHub: github.com/rudgebreen/DocumentingTransmission–The-Rake-Cycle/. Accessed: 26 Apr. 2019].

———. *Documenting Transmission: Digitally Tracing and Displaying the Folk Process*. M.phil. thesis, Maynooth University, 2016. www.academia.edu/34643263/Documenting_Transmission_Digitally_Tracing_a. Accessed 26 Apr. 2019.

Burns, Robert G.H. "Continuity, Variation, and Authenticity in the English Folk-Rock Movement." *Folk Music Journal*, vol. 9, no. 2, 2007, pp. 192–218. *JSTOR*, www.jstor.org/stable/4522807. Accessed 30 Mar. 2018.

Clement, Tanya. *In Transition: Selected Poems by the Baroness Elsa von Freytag-Loringhoven.* 19 May 2016. digital.lib.umd.edu/transition/source. Accessed 30 Mar. 2017.

———. "Knowledge Representation and Digital Scholarly Editions in Theory and Practice." *Journal of the Text Encoding Initiative*, issue 1, 2011. jtei.revues.org/203. Accessed 12 Nov. 2017.

———, editor. *In Transition: Selected Poems by the Baroness Elsa von Freytag-Loringhoven.* digital.lib.umd.edu/transition. Accessed 30 Mar. 2017.

Cohen, N. "Folk music." *Grove Music Online*, 2016. www.oxfordmusiconline.com/subscriber-/article/grove/music/A2241135. Accessed 30 Mar. 2017.

Comhaltas Traditional Music Archive, 2000–2009. archive.comhaltas.ie. Accessed 21 Aug. 2019.

Cowdery, James R. "Kategorie or Wertidee? The early years of the International Folk Music Council." *Répertoire International de Littérature Musicale.* www.rilm.org/historiography/-cowdery.pdf. Accessed 12 Nov. 2017.

Goldstein, Kenneth S. *The Unfortunate Rake: A Study in the Evolution of a Ballad*, liner notes and musical recording, Smithsonian Folkways, 1960.

———. "Still More of "The Unfortunate Rake" and His Family." *Western Folklore*, vol. 18, no. 1, 1959, pp. 35–8.

Grier, James. "The Critical Editing of Music." *The Critical Editing of Music: History, Method, and Practice*, edited by James Grier, Cambridge University Press, 1996, pp. 10–3.

Hand, Wayland D. "The Cowboy's Lament." *Western Folklore*, vol. 17, no. 3, 1958, pp. 200–5. *JSTOR*, www.jstor.org/stable/1496046. Accessed 30 Mar. 2017.

Hankinson, Andrew, et al. "The Music Encoding Initiative as a Document-Encoding Network." *12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, 2011. www.mirlab.org/conference_papers/International_Conference/ISMIR%202011/papers/OS3-1.pdf. Accessed 12 Nov. 2017.

Harwood, Robert W. *I Went Down to St. James Infirmary.* Harland Press, 2008.

Kerman, Joseph. "Sketch Studies." *19th-Century Music*, vol. 6, no. 2, 1982, pp. 174–80. JSTOR, www.jstor.org/stable/746277. Accessed 30 Mar. 2018.

Knight Lab. *StoryMapJS.* storymap.knightlab.com/advanced. Accessed 30 Mar. 2017.

Lodewick, Kenneth. "The Unfortunate Rake" and His Descendants. *Western Folklore*, vol. 14, no. 2, 1955, pp. 98–109. JSTOR, www.jstor.org/stable/1496993. Accessed 30 Mar. 2018.

The Music Encoding Initiative (MEI). music-encoding.org. Accessed 30 Mar. 2017.

MusicXML, www.musicxml.com. Accessed 30 Mar. 2017.

The Online Chopin Variorum Edition (OCVE). www.chopinonline.ac.uk/ocve. Accessed 30 Mar. 2017.

Propp, V. *Morphology of the Folk Tale.* The American Folklore Society, 1968.

Roland, P. "The Music Encoding Initiative (MEI) DTD and the OCVE." *Online Chopin Variorum Edition.* 30 Dec. 2015, pilot.ocve.org.uk/redist/pdf/roland.pdf. Accessed 30 Mar. 2017.

Vaughan Williams Memorial Library (VWML). *Roud Folk Song Index.* www.vwml.org/. Accessed 30 Mar. 2017.

Schreibman, Susan. "Re-Envisioning Versioning: A Scholar's Toolkit. *Digital philology and medieval texts*, edited by Arianna Ciula and Francesco Stella, " 2007. www.infotext.unisi.it/upload/DIGIMED06/book/schreibman.pdf. Accessed 30 Mar. 2017.

Schreibman, Susan, et al. "The Versioning Machine." *Literary and Linguistic Computing*, vol. 18, no. 1, 2003, pp. 101–7.

———, et al. *Versioning Machine 5.0.* 2016. v-machine.org/. Accessed 30 Mar. 2017.

———, editor. *The Thomas MacGreevy Archive.* 2007. www.macgreevy.org/. Accessed 21 Aug. 2019.

Sharp, Cecil J. *English Folk Song, Some Conclusions.* Simpkin & Co, London, 1907. archive.org/-details/englishfolksongs00shar. Accessed 30 Mar. 2017.

Sperberg-McQueen, C.M. "Textual Criticism and the Text Encoding Initiative." *MLA*, 1994. www.tei-c.org/Vault/XX/mla94.html. Acessed 30 Mar. 2017.

TEI Consortium, editor. *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.4.0.* 2018, www.tei-c.org/Vault/P5/3.4.0/doc/tei-p5-doc/en/html.

TEI Roma: generating customizations for the TEI. www.tei-c.org/Roma/. Accessed 30 Mar. 2017.

Thompson, Stith. *Motif-Index of Folk-Literature.* Indiana University Press, Bloomington, 1955–58. sites.ualberta.ca/ urban/Projects/English/Motif_Index.htm. Accessed 30 Mar. 2017.

Washbourne, Kelly. "Translation, the 'Folk Process', and Socially Committed Songs of the 1960s. "*Mutatis Mutandis: Revista Latinoamericana de Traducción*, vol. 6, no. 2, 2013, pp. 455–76.

You Can't Put Your Arms Around a Memory—The Multiple Versions of Alexander von Humboldt's "Kosmos-Lectures"[1]

Christian Thomas

Abstract

I present a collection of primary sources related to Alexander von Humboldt's world-famous "Kosmos-Lectures." These lectures, held in Berlin in 1827/28, mark a milestone in the history of sciences and their popularization. Given their indisputable significance, surprisingly little research has been conducted on the lectures. One reason for this was, until recently, the lack of primary sources available. With the online-publication of all currently known lecture notes by attendees of the "Kosmos-Lectures" and the digitisation of Humboldt's legacy collection, this situation has changed significantly: While before we had too few, it now seems as if we had too many witnesses—or *versions*—of Humboldt's lectures. I argue that each document represents one distinct, equally valid version of the "Kosmos-Lectures" that has to be presented and appreciated in its own right. Even if we had the most intimate sources at hand, it would still be impossible to reconstruct an event like a public oral lecture. We remain struck with a multitude of witnesses, i.e. *versions* at hand. I believe that the implications derived from this exemplary corpus are transferrable to many other, similar instances in which we necessarily are dealing with various, but always "indirect" historical transmissions.

[1] This paper presents an argument first made at the Versioning Cultural Objects Symposium, held in December 2016 at Maynooth University (Ireland). The text has been revised following very instructive and valuable feedback by other participants of the symposium, and again after gratefully receiving important suggestions from peer reviewers and the editors of this volume until the end of the year 2018. In the meantime, until the end of 2019, we have discovered some more details in the context of Alexander von Humboldt's Kosmos-Lectures. Among these are the identification of Henriette Kohlrausch as the scribe of the currently only known transcript of Humboldt's lectures at the Sing-Academy hall, referenced here as N.N. a, and the publication of said notebook in printed form. Please see the extensive foreword to this edition by Christian Thomas and Christian Kassung (*Humboldt/Kohlrausch* 9–58), on the current state of research. Additional to this volume, a lot more has been published by and about Alexander von Humboldt in years since 2018, and especially in the context of the international celebrations of his 250th anniversary in 2019. However, these new findings and more recent publications do not alter the main thesis of the paper at hand, and therefore the remainder of the text has not been updated in detail. — Christian Thomas, Berlin, September 2019.

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 77–99.

1 Humboldt's "Kosmos-Lectures" (1827/28): a hot spot in the history of sciences, a blind spot of research

Soon after his definite return from Paris to Berlin in May 1827, the Prussian-born naturalist Alexander von Humboldt (1769–1859) announced a series of lectures on "Physical Geography" in the wider sense that the subject encompassed in his understanding. In these so-called Kosmos-Lectures, held in Berlin from fall 1827 until April 1828, Humboldt presented the scientific knowledge of his time, covering an extraordinary range of natural phenomena and scientific disciplines. He held two separate public courses: at the Berlin University, an unprecedented number of about four hundred students, professors, members of the court, and private scholars gathered for a total of sixty-two lessons. Soon after starting this first course, Humboldt opened up a parallel, second series of sixteen lectures on the same topics at the nearby Sing-Academy building. For this public course a larger, more diverse crowd of around one thousand people gathered.

When reflecting upon the Kosmos-Lectures in general, and especially in the context of this paper, where records of the lecture courses penned by attendees play an important role, it is crucial to get an idea of the audience Humboldt was addressing. Although there is no existing list of attendees for either set of lectures that we know of, we can infer from contextual sources that it included students and professors, members of the Prussian court, and also interested laymen. The course at the Berlin University was announced as "öffentlich" ("publice"), which, in this time and context, also indicates that the lessons were admissible free of charge.[2] To make sure that the broadest possible public had access to the knowledge he presented was an important point on Humboldt's agenda:[3] By paying the rent and the costs for heating for the Great Lecture Hall at the Sing-Academy building out of his own pocket, Humboldt made sure that this course could likewise be attended without an entry fee.[4] He also

---

[2] Quotations from *Verzeichniß* 6. On the notions of "publice" vs. "privatim", see *Die Vorlesungen der Berliner Universität 1810–1834* XVII f. Contemporaries appreciated Humboldt's decision, which was not a matter of course, and demanded that more academics should follow his lead and ensure cost-free public access to education: "[d]ie anderen begüterten Professoren sollten auch so edel seyn, *öffentlich* vorzutragen, und sich die Wissenschaft *nicht schwer bezahlen lassen*."(*Außerordentliche Beilage zur Allgemeinen Zeitung 1827*, Nr. 41, p. [161]; emphasis mine.). Single articles from newspapers, archival material from Humboldt's legacy collection, and webpages mentioned here will not be listed in the bibliography section of this paper, but referenced only in the footnotes.

[3] He vehemently protested when an article in the international newspaper Moniteur Universel claimed that, on the contrary, he was taking money from his listeners. See, for example *Le Moniteur Universel*, No. 66, Jeudi, 6 Mars 1828; *Neue Zürcher-Zeitung*, No. 22, March 15, 1828; *Allgemeine Zeitung*, No. 81, March 21, 1828.

[4] See, for example *Oesterreichischer Beobachter*, March 2nd, 1828, p. 252 (quoting the *Preußische Staatszeitung* from February 23rd, 1828): "Weit entfernt, den Zutritt zu seinen Vorlesungen durch die Erlegung irgend eines Honorars zu bedingen, darf H$^{r.}$ v. Humboldt ganz besonders in dem zweiten Cursus die

emphasised that women were invited to attend the second set of lectures, even though they were excluded from Prussian Universities until the end of the 19th century.

The range of topics, the genius of the lecturer, the number and diversity of the participants, and the lasting impression this event made on the public and governmental participants was exceptional. The Kosmos-Lectures indisputably mark an important milestone in the history of sciences and in the genesis of the concepts and methods central to their rise in the 19th century.[5] Against this backdrop, it is surprising how little research has been conducted on the lectures both in terms of the actual content and also regarding the essential differences between the two separate courses. Neither has their relationship to contemporary and later publications by Humboldt himself been examined, nor have they been investigated alongside the works of other scientists of the time. In this regard, the famous Kosmos-Lectures still remain an under-researched topic.[6]

The central argument presented in this paper is that the main reason for this observable absence of research on the lectures is the lack (or, with the same result, the neglect) of witnesses documenting the event itself. But since 2016, with the online publication of all currently known individual notebooks written by attendees of the Kosmos-Lectures[7] and with the digitisation of Humboldt's legacy collection held in Krakow and Berlin,[8] this situation has changed considerably. Until recently, there were too few documents to base substantial research on; now, with the availability of Humboldt's (fragmentary) manuscripts, several of his listener's accounts and other related documents available online, it seems as if we now have too many witnesses—or *versions* —documenting Humboldt's oral presentation.

In the remainder of this paper, I will give a more detailed overview of the current state of research concerning Humboldt's Kosmos-Lectures, focusing on the question of which primary sources have been known and were readily available as the

---

Zuhörer als seine Gäste betrachten, da die nicht unbeträchtliche Ausgabe für Miethe und Heitzung des Saales ihm allein anheimfällt."

5 For a very popular and recent publication see Wulf 193–196, which sums up in short the significance assigned to the lectures in general. As inaccurate as this condensed passage is in some details, it gives a good impression of what can be considered common knowledge regarding the Kosmos-Lectures among academics as well as the wider public. On Humboldt's contribution to the enforcement of scientific methods and research in Prussia compare, for example, Klein.

6 See Erdmann and Thomas for a more detailed overview on the state of research until that time.

7 See Humboldt Universität zu Berlin: *Hidden Kosmos*, 2014–2016, www.culture.hu-berlin.de/hidden-kosmos. Accessed 2 Dec. 2018. For a detailed list of all currently known notebooks see www.culture.hu-berlin.de/de/forschung/projekte/hidden-kosmos/veroeffentlichte-nachschriften. Accessed 2 Dec. 2018, and Thomas, et al. on the project's principle aims and methods. The full text transcriptions are published by our cooperation partner *Deutsches Textarchiv* at the Berlin-Brandenburg Academy of Sciences and Humanities, www.deutschestextarchiv.de/. Accessed 2 Dec. 2018.

8 See humboldt.staatsbibliothek-berlin.de/werk/, sections "Nachlass Alexander von Humboldts in Berlin" and "Nachlass Alexander von Humboldts in Krakau." Accessed 2 Dec. 2018.

documentary foundation of such research. The next section will illustrate the shift from a very sparse documentary base to a wealth of witnesses due to the recent digitisation of important primary sources. I will then focus on the question of how to deal with this newfound wealth from a methodological point of view. Furthermore, I will discuss the extent to which this methodological approach to dealing with competing/complementing versions affects the practical side of working with these primary materials.

2 Primary sources on Humboldt's Kosmos-Lectures: from drought to deluge

The observable lack of research focusing on the Kosmos-Lectures directly leads to questions concerning the material base on which research could have been built upon. In several respects, the balance is sobering: there is no printed publication of the lectures authorized by Humboldt himself.[9] His original script, as will be elaborated in more detail in the remainder of this paper, is not preserved in its initial form. Over the years that followed the 1827/28 lectures, Humboldt altered the manuscript he had used significantly by rephrasing, updating, reordering, and partly discarding his former lecture notes. What is left needs to be sorted and examined meticulously.[10] Fortunately, several of his hundreds of listeners made extensive notes covering the lectures. The keeping of notebooks was, as many examples prove, a widespread practice of this time, especially in the academic context.[11] However, only two of these notebooks were available in printed editions (*Alexander von Humboldts Vorlesungen*, printed in 1934 and Humboldt's *Über das Universum* from 1993), wheras the majority of these pivotal primary sources to the lectures, held in different libraries and private collections, remained practically unknown. Until recently, they had not been listed systematically, let alone been transcribed and published in an edition.

---

[9] Although a contract concerning the publication of the lectures was sealed between the author and his publisher Cotta in March 1828 (i.e. when the lectures were still in progress) Humboldt's later publication, *Kosmos* (1845–62), must not be considered as simply an elaborated, now printed version of the former lectures. See Werner on the genesis of the *Kosmos* and its undisputed status as an original publication.

[10] Before the digitisation of Humboldt's legacy collection and the accompanying archival indexing of thematic provenance in 2016, this task would have been a tedious, on-site labour. Now, at least the fragments of the original manuscripts are accessible online, but the task of identifying the scattered parts among the vast collection Humboldt left as his scientific legacy remains an outstanding, challenging task, as will be explained later in this section.

[11] Compare, to name just, to name just, to name just a few prominent examples, the numerous editions of famous lectures by Kant, Hegel, Schelling, Lichtenberg, Nietzsche, Schleiermacher, and others, that are in part or even entirely derived from one or several attendee's notebooks. See also the contemporary handbook Fischer dedicated to that practise in 1826, offering students and private scholars advice on how to keep a notebook of an academic or educational lecture.

The lack of research on the Kosmos-Lectures can therefore be explained by the sparse documentary base of (commonly known and easily accessible) primary sources for any investigation until recently. Another important factor is the widespread conception that the lectures represent merely a stepping stone in Humboldt's "publication biography," becoming more or less obsolete with the appearance of Humboldt's last and probably most famous publication, *Kosmos. Entwurf einer physischen Weltbeschreibung*.[12] From this perspective, the Kosmos-Lectures are only an early attempt to lay out contemporary scientific knowledge in a holistic, descriptive and still aesthetically appealing "Naturgemälde" ("portrait of nature"); an attempt which is then superceded in every respect by the five volumes of the *Kosmos*.[13] As a consequence, the preceding Berlin lectures are not considered an independent (oral) publication, and may have appeared to be less interesting and less vital as a subject of study in themselves.

While the *Kosmos* remained fragmentary, and important topics such as the geography of plants or the ethnic diversity and geographical distribution of man—as well as other topics—were not elaborated upon in due detail, each cycle of the Kosmos-Lectures was complete in itself. The sixty-two-lesson University course as well as the sixteen-lesson Sing-Academy course encompassed a full "panoramic" view, starting from the astronomical and proceding via the terrestrial to intelligible and cultural phenomena, respectively. Additionally, the Kosmos-Lectures are connected to and—as will be demonstrated with examples in the remainder of the paper—contain set pieces of several publications other than the *Kosmos*. It is therefore crucial for further research to overcome the narrow view that the Kosmos-Lectures are a more or less negligible step on the way to the printed *Kosmos*—as well as the fixation on the latter as the former's principle point of reference.

Another part of the reason why the Kosmos-Lectures have been neglected as an object of study can be found in the in the *Kosmos*' first volume from 1845: in the foreword, Humboldt evokes the distant, yet cherished memory of his public lectures held in Berlin in 1827/28 (and, in the years before that, in Paris). The strategy behind this was twofold, with rather conflicting, if not mutually exclusive ends: on the one hand, Humboldt wanted to build upon their success. He wished for his printed publication to reach a wide audience, and the popularity of the past lectures to help in finding a broad readership for the *Kosmos*. On the other hand, he wanted to rule out the assumption that the *Kosmos* was based on papers from way back then, i.e. material that would be outdated. For this reason, Humboldt makes the rather surprising claim:

---

[12] *Kosmos* was immediately translated (at first without approval of the author) into French and English, and subsequently into almost every other major language.

[13] The foundation for this perspective was laid in the first encompassing, scientific biography on Humboldt, in the section where Alfred Dove describes Humboldt's Berlin years (*Alexander von Humboldt: Eine wissenschaftliche Biographie*, II, 137–40).

> Bei freier Rede habe ich in Frankreich und Deutschland *nichts über meine Vorträge schriftlich aufgezeichnet.* Auch die Hefte, welche durch den Fleiß aufmerksamer Zuhörer entstanden sind, blieben mir unbekannt, und wurden daher bei dem jetzt erscheinenden Buche auf keine Weise benutzt. (*Kosmos. Entwurf*, I, X)

He hastens to add that the current work is a recent creation and that the similarities between the past lectures and the current print are restricted to the conceptual level.[14] Especially the first statement, i.e. the claim that Humboldt had never written down anything in preparation for his public lectures, seems too unlikely to be true. To imagine that Humboldt presented a total of sixty-two lessons at the University plus an additional, parallel sixteen lessons at the Sing-Academy on highly complex matters without a script, i.e. improvising each lesson or recounting it from memory, is hard to believe. As one would suspect, the claim is demonstrably false, as both contemporary witnesses and as Humboldt's first biographer Alfred Dove testified to decades ago.[15] But for many years Humboldt's claim was obviously taken at face value, which led to the neglect of important documents by the research community, among these the surviving remains of Humboldt's original manuscript.

The second part of the statement cited above is interesting because Humboldt mentions another possible source to the contents of the past lectures: the notebooks kept by his "the industry of certain attentive auditors." Humboldt, of course, was well aware of their existence: while the Kosmos-Lectures were still in progress, he had forbidden any publication of such notebooks,[16] not only to assure his prerogative right

---

[14] For an English translation, see *Cosmos. A Sketch*, xii: "My lectures were given extemporaneously, both in French and in German, and without the aid of written notes, nor have I, in any way, made use, in the present work, of these portions of my discourses which have been preserved by the industry of certain attentive auditors." One reason for this "deception strategy" probably was that the lectures from the 1820s were long past when the first *Kosmos*-volume finally appeared in 1845. Humboldt obviously wanted to make sure that no reader would assume that the content presented in print might be outdated, therefore asserting, "Die Vorlesungen und der Kosmos haben also nichts mit einander gemein als etwa die Reihefolge der Gegenstände, die sie behandelt" ("[…] my lectures and the Cosmos have nothing in common beyond the succession in which the various facts are treated."). (Ibid.)

[15] See *Alexander von Humboldt: Eine wissenschaftliche Biographie*, II, 137. Probably the earliest publication bearing evidence that Humboldt had an elaborated script prepared for each lecture is the correspondence between Humboldt and Karl August Varnhagen von Ense (1785–1858), first published in 1860. Following complaints from G. W. F. Hegel (1770–1831), who had head (from his—Hegel's—followers) that Humboldt had attacked his (Hegel's) philosophy of nature in one of the lectures, Humboldt asked his confidante, Varnhagen, to mediate a peace, and forwarded the manuscript of the lecture in question to be passed on (see *Briefe von Alexander von Humboldt an Varnhagen* 3). Humboldt's notes must have been rather extensive in order to serve as proof that he indeed did not attack Hegel. Another clue to their elaborated status is that Humboldt asks Varnhagen to make any use of the papers, except to duplicate them for publication. This addition would have been unnecessary if the sheets had contained only a few keywords.

[16] *"Spenersche" Zeitung*, Dec. 12, 1827, p. [7]; paraphrased in *Neckar-Zeitung*, No. 356, Dec. 29, 1827,

to their (print) publication, but also because he mistrusted their quality as sources.[17] He knew that each attendee would necessarily come up with his or her individual perspective, an incomplete and somewhat distorted account of what was said. We can safely assume that Humboldt did not use any of these while working on the *Kosmos*: first, because he did not value these sources very highly and second, because (contrary to what he said) he had his own, more accurate papers to return to.

At the end of this introductory section, we have an idea of the most important groups of primary sources on Humboldt's Kosmos-Lectures: the lecturer's own manuscripts and the notebooks of his attendees. A third group is the material that Humboldt used to prepare his manuscripts, and includes previous publications of his, letters from other scholars he corresponded with, and articles and monographs published by his colleagues. The concept of *versions* is as important for the presentation of these primary sources as it is for their interpretation. In the remainder of this paper, I focus on three classes of *versions* of Humboldt's Kosmos-Lectures:

1. (ideally all of) the original notes that Humboldt used to deliver his lessons (which he has revised intensively and reorganised in the years following the lectures);
2. (ideally all) notes taken by his auditors (which offer a great deal of variety among each other);
3. (ideally all) material used by Humboldt to prepare his lessons, e.g. preceding publications of his own and other researchers, letters and excerpts, etc.

Considering the number of primary sources related to the lectures and the complex relations between these documents, I consider these documents to be a set of multiple versions of Humboldt's Kosmos-Lectures. These versions have different authorial statuses: they sometimes complement each other, sometimes run parallel, and sometimes contradict each other. I argue that this irritating polyphony is inherent to the qualities of our research object: like a distant memory—that, as Johnny Thunders put it, "you can't put your arms around"—the event of the lectures as a singular performance eludes our grasp. It cannot be repeated or reconstructed in a definitive shape, but can only be recounted from different, equally limited perspectives.

---

p. 1641.

[17] See also a later letter from Humboldt to Richard Zeune (1817–1875), Berlin, Feb. 16, 1857, where Humboldt unmistakeably states his aversion to (the publication of) these documents: "[…] nichts ist widerwärtiger, als publicirt zu sehen, was ein Gemisch von Gehörtem und Selbstzugesetztem ist." ("Nothing is more repugnant than to see publicised what is a mixture of what is heard and what is self-imposed." Quoted from *Alexander von Humboldt: Eine wissenschaftliche Biographie*, II, 137; translated by Christian Thomas).

3 Humboldt's original lecture manuscript: A dismembered corpus of prewritings and rewritings

Since Humboldt himself never published the lectures as such, his original script would seem to be a natural candidate for the most reliable and complete source: an authoritative version of what was most likely said at the lectern. As stated above, Humboldt later claimed that he had spoken extemporaneously the entire time and had no preparatory notes. This claim had long been falsified by the reliable accounts by several contemporary eyewitnesses[18] and there are a considerable—but as of yet unknown—number of sheets related to the Kosmos-Lectures preserved in Humboldt's legacy collection that he clearly used for both lecture series. But since his denial of their existence apparently was taken as a fact by researchers, Humboldt's lecture notes as a whole still remain largely unknown.

Since the end of 2016, Humboldt's complete papers can be accessed in digital form via the Berlin State Library, where the entire collection has been digitised and was virtually reunited with the parts that have remained in the Jagiellonian Library in Krakow after the Second World War ended.[19] Among these papers are several sheets that evidently have been used for the Kosmos-Lectures (see fig. 1).

In the process of digitisation, the material from Humboldt's legacy collection—which is for the most parts still preserved in its original order and had been catalogued after Humboldt's death in 1859—was thoroughly re-examined and furnished with additional metadata. In the process, the documents were assigned to a context of usage whenever possible, e.g. the thematic collection or publication project Humboldt used the material in question for is given in the archival metadata record.[21] The granularity of the metadata records varies greatly, depending on the physical structure

---

[18] Several accounts by auditors of the lectures suggest that Humboldt not only used his notes as a guidance while freely extemporising, but even read out loud whole passages, which for example his niece, Gabriele von Bülow, (1802–1887) found "not pleasant"(*Gabriele von Bülow – Tochter* 195). The unusually high degree of similarity between certain passages from the auditors' notebooks and published articles by Humboldt indicates that on several occasions, Humboldt was reading out written material out word-by-word.

[19] See the project's website humboldt.staatsbibliothek-berlin.de; and Erdmann and Weber. The legacy collection was digitised along with the recently acquired American travel journals that are currently being edited in a hybrid edition at the Berlin-Brandenburg Academy of Sciences and Humanities by the Academy Project "Alexander von Humboldt auf Reisen – Wissenschaft aus der Bewegung," www.bbaw.de/en/research/avh-r. All accessed 2 Dec. 2018.

[20] Examples taken, in order of appearance, from SBB-PK, Nachl. A. v. Humboldt, kl. Kasten 3b, Nr. 73, p. [11], resolver.staatsbibliothek-berlin.de/SBB00018C3600000011. Accessed 2 Dec. 2018; gr. Kasten 11, Nr. 16, p. [1], resolver.staatsbibliothek-berlin.de/SBB0001AB9300000001. Accessed 2 Dec. 2018; gr. Kasten 8, Nr. 5a, Bl. 3r, resolver.staatsbibliothek-berlin.de/SBB0001676C00000005. Accessed 2 Dec. 2018.

[21] The "Findbuch" for the Nachlass Alexander von Humboldt gives a good idea of its structure and extent, see Kalliope Portal, "Online-Ansicht des Findbuchs Nachl. Alexander von Humboldt," kalliope-verbund.info/de/findingaid?fa.id=DE-611-BF-4430. Accessed 2 Dec. 2018.

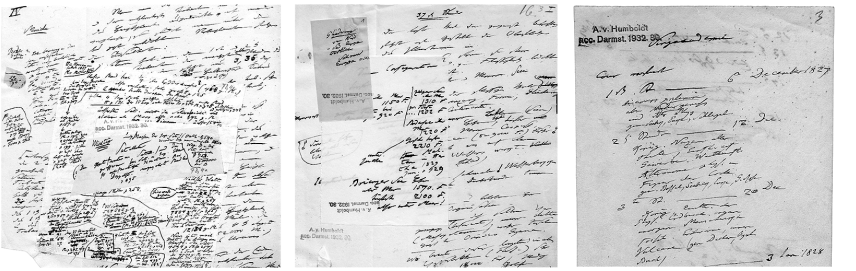


Figure 1: Manuscripts related to the Kosmos-Lectures from Humboldt's legacy collection at the Berlin State Library. *Left and center*: revisions, notes and several layers of smaller sheets with additions attached to the original notes; *on the right*: detail of Humboldt's outline of the Sing-Academy course.[20]

in which Humboldt kept his notes and on the number of documents he had filed under one thematic complex: sometimes there are hundreds of documents—including manuscripts in Humboldt's hand, letters, fragments of print publications, datasheets, tables, and lists provided by third parties, etc.—grouped together in one envelope and therefore recorded as one archival item containing hundreds of numbers; sometimes there are only a few sheets forming one record.

Discovering certain items, in this case documents belonging to the Kosmos-Lectures, can thus be a difficult task, especially since Humboldt continued to work with the papers after the lectures, and redistributed them across his vast collection of working material. The whole collection follows a thematic order rather than a chronological or project-oriented order, and following this logic, the different parts of the lecture manuscript were redistributed among the entire collection and ended up in the company of documents of different origin and different (initial) purpose. As a result, Humboldt left no coherent, closed set of documents explicitly labelled "Kosmos-Lectures." We also have to assume that the original notes were preserved only fragmentarily, since Humboldt seems to have discarded material that he considered outdated or less relevant by the time of their re-examination years after the lectures. Making matters more complicated, Humboldt, in the course of reorganising the remaining material, constantly added supplementary information and new findings over the years following the lectures, thereby altering, revising, and partly overwriting the original text base. The resulting "bricolages," some of which can be seen in figures 1–3, are an amazing example of an analogue, material database.

Humboldt's decade-spanning labour of reordering, altering, and supplementing the notes initially prepared for the Kosmos-Lectures imposes a further obstacle when trying to positively identify those papers that originally belonged to the lecture's

manuscript. However, some of these manuscripts can quite easily be assigned to the Kosmos-Lectures, especially those where Humboldt noted the particular lesson to which they belonged, as the example in figure 2 illustrates. On the top of the page in the background, Bl. 85r, which is partly covered by smaller sheets attached to it, Humboldt wrote "52te Stund[e]" (52nd lesson) and, as thematic keywords, "Electrische Erscheinungen" ("Electrical Phenomena"); Bl. 87r contains the same note "52te Stund[e]." This manuscript (or at least parts of it) therefore must have initially belonged to the Kosmos-Lectures, more specifically to the course at the Berlin University, where Humboldt held sixty-two lessons, while the Sing-Academy course ended after only sixteen lessons.[22]

Humboldt attached additional sheets containing notes and bibliographic references by using fixation dots made of wax and glue. Thus, the sheets can be moved to the side, uncovering the underlying text without loss (see figure 3). This technique of fixing single notes by gluing them together, at the same time preserving their mobility and hindering text loss is typical—for Humboldt's work as it is manifested in his legacy collection. But unfortunately hints like the text "52te Stund[e]" in our example, which clearly assigns the base leaf and a second part of the collage to a certain lesson of the Kosmos-Lectures, are not typical at all: Only some parts of the original manuscript contain such an explicit reference to a certain part of the lessons[23]; others can only be recognized from context. Altogether, these documents are hard to identify, especially since Humboldt dissolved their original order and succession as described above. In this important respect, the manuscript in our example is also typical, as it shows why Humboldt's lecture notes, even if they were preserved in its entirety, could not be used to reconstruct the original contents of the Kosmos-Lectures. Two of the sheets in the foreground of fig. 2, Bl. 88r and 89r, contain bibliographical references from 1829 and 1834, and can therefore only have been attached after the completion of the lectures in 1828. While it is thus clear that these were not part of the original manuscript, the same is also possible, but harder to determine, for the other sheets.

---

[22] The assumption that the reference "52te Stund[e]" and other, similar references of that kind point to the respective lesson of the Kosmos-Lectures (and not to some other context) was confirmed by comparing the contents and keywords to the attendee's notebooks.

[23] For other manuscripts containing references like this, see, for example:
Nachl. Alexander von Humboldt, gr. Kasten 12, Nr. 142, Bl. 9r (4th lesson),
resolver.staatsbibliothek-berlin.de/SBB0001A5C300000025;
gr. Kasten 11, Nr. 16, Bl. 3r (37th lesson),
resolver.staatsbibliothek-berlin.de/SBB0001AB9300000007;
gr. Kasten 11, Nr. 19a, Bl. 2r (39th lesson),
resolver.staatsbibliothek-berlin.de/SBB0001AB9700000005;
gr. Kasten 13, Nr. 15, Bl. 76r (59th lesson),
resolver.staatsbibliothek-berlin.de/SBB0001B83200000175;
gr. Kasten 13, Nr. 15, Bl. 15r (60th lesson),
resolver.staatsbibliothek-berlin.de/SBB0001B83200000036. All accessed 2 Dec. 2018.

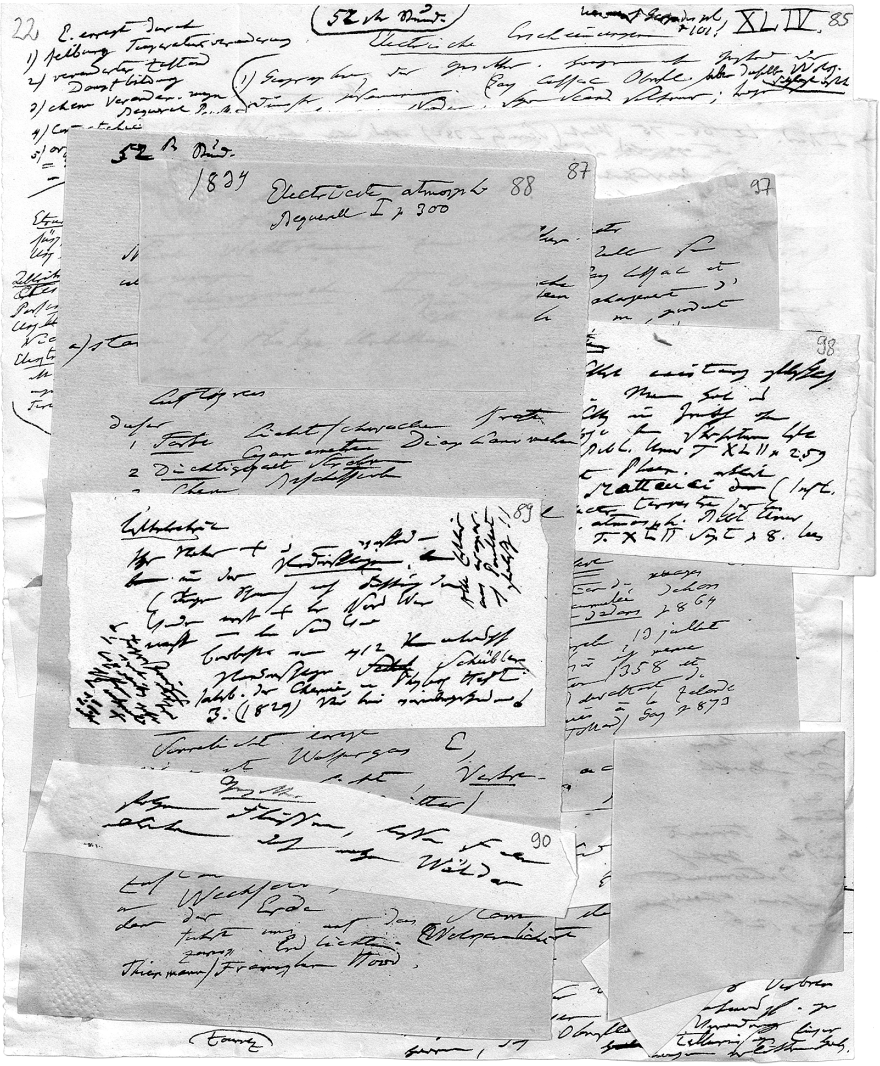


Figure 2: Nachl. Alexander von Humboldt, gr. Kasten 12, Nr. 16, Bl. 85r–98r; several smaller sheets attached on top of one base leaf, resolver.staatsbibliothek-berlin.de/SBB0001A52B00000205. Accessed 2 Dec. 2018.

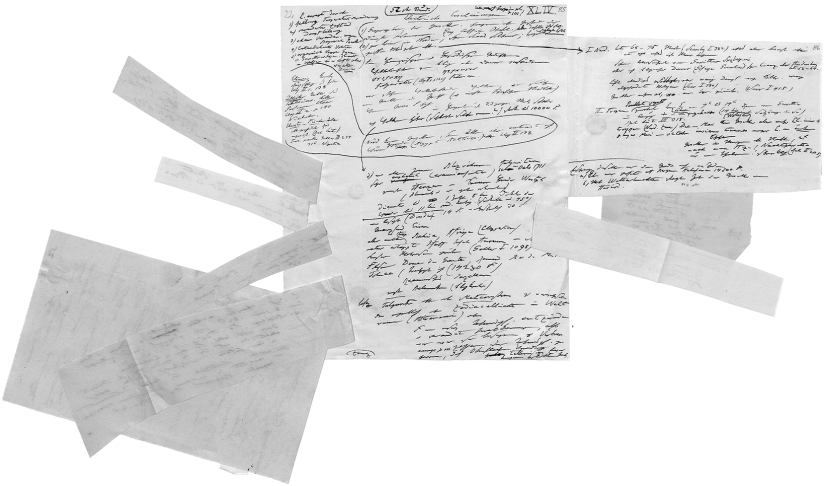


Figure 3: Smaller sheets unfolded, revealing the text from Bl. 85r and 86r covered in fig. 2; Nachl. Alexander von Humboldt, gr. Kasten 12, Nr. 16, Bl. 85r–98r; resolver.staatsbibliothek-berlin.de/SBB0001A52B00000209. Accessed 2 Dec. 2018.

The connection and succession of the original lecture manuscripts is destroyed and may not be reconstructed in total. In this respect, the situation is not as ideal as initially imagined: Humboldt's original manuscript is not complete anymore; it has been re-organized, transformed and altered significantly over many years. As unfortunate as this might seem, one might find comfort in the central assumption I defend in this paper: that is, even if we did have each and every single page of some completely elaborated papers in its original, contemporary state, we could never determine whether Humboldt stuck to the script, or if he was distracted in his flow of words by some objection, led astray by a random observation that day, etc. Therefore, even the most comprehensive lecture scripts would not give proof of what was actually uttered. Like every other primary source I present here, they would offer merely one version among possibly many others.

4 The auditors' notebooks: quotations, paraphrases, and misrepresentations

Another important source for the lectures are the above-mentioned handwritten notebooks by Humboldt's auditors. Humboldt himself was well aware of their existence and (not unrealistically, considering the number of attendees) assumed their

number to be in the hundreds. But he detested their inherent flaws and inevitable inaccuracies and even interdicted their publication.[24] This may be one of the reasons why it wasn't until 1934, 106 years after the lectures, until the first auditor's notebook from the University class was published in a printed edition (*Alexander von Humboldts Verlesungen*). More recently, in 1993, the edition of a second notebook, this one covering the Sing-Academy lectures, followed (Humboldt, *Über das Universum*). Until the end of 2014, only these two were available as full-text transcriptions in printed form. Unfortunately, both editions are scientifically inadequate: they do not meet standards of scholarly editing and contain many transcription errors. The existence of at least five other attendee's notebook was known for decades,[25] but they were never edited and were only accessible in the handwritten original. Since working with manuscripts is a tedious task and the witnesses were held in different places, these additional archival sources remained practically unknown.

Fortunately, by the end of 2016, with the conclusion of the two-year *Hidden Kosmos*-project funded by the Excellence Initiative at Berlin's Humboldt University,[26] and the publication of another, formerly unknown (fragmentary) notebook—which the author of this paper discovered only in March 2017—as an addition to the *Hidden Kosmos*-corpus,[27] this situation has changed significantly. Currently, we know of twelve manuscripts altogether, nine of which are related to the University and three to the Sing-Academy lectures (figure 4).

Eleven of these manuscripts were published as full text transcriptions encoded in TEI-XML[29] via the *Deutsches Textarchiv (German Text Archive)*[30] in 2016. Among

---

[24] See, as one example of the many venues that printed and reprinted Humboldt's statement, the *"Spenersche" Zeitung*, Dec. 12, 1827, p. [7]: "Obgleich ich der Besorgniß nicht Raum geben möchte, daß Hefte, welche Zuhörer meiner Vorlesungen zu ihrer Erinnerung schreiben, durch Zufall in andere Hände kommen und gedruckt werden könnten, so halte ich es dennoch für besser, hierdurch öffentlich zu erklären, daß ich jede Publikation dieser Art, als einen Eingriff in mein Eigenthum betrachten werde." i.e., Humboldt basically states that he will consider each publication of such notebooks as an interference with his (intellectual) property, i.e. as an instance of what would today be considered a copyright violation.

[25] See Engelmann 28, where seven notebooks are mentioned, six of which were still unpublished at that time.

[26] See also footnote 7 of this paper.

[27] See Willisen. The fortuitous discovery of this notebook in a legacy collection in the Geheimes Staatsarchiv Preußischer Kulturbesitz proves that there are still more notebooks to be found.

[28] Except for items 4), 7) and 10), where the holding institution and person are stated separately, the manuscripts are held at the State Library in Berlin, and were, except for item 8), published by the *Hidden Kosmos*-Project via *Deutsches Textarchiv*: 1) Parthey; 2) [N.N.] c; 3) Riess; 4) Libelt; 5) Patzig; 6) [N.N.] b; 7) [N.N.] d; 8) Lohde; 9) [N.N.] a; 10) Hufeland.

[29] I.e. formatted in the platform-independent Extensible Markup Language (XML) following the Guidelines of the international *Text Encoding Initiative* (TEI). See the *TEI* website, www.tei-c.org/index.xml; and the P5 Guidelines, http://www.tei-c.org/Guidelines/P5/. Both accessed 2 Dec. 2018.

[30] The encoding follows the recently developed DTA Base Format for Manuscripts (DTABf-M), a true

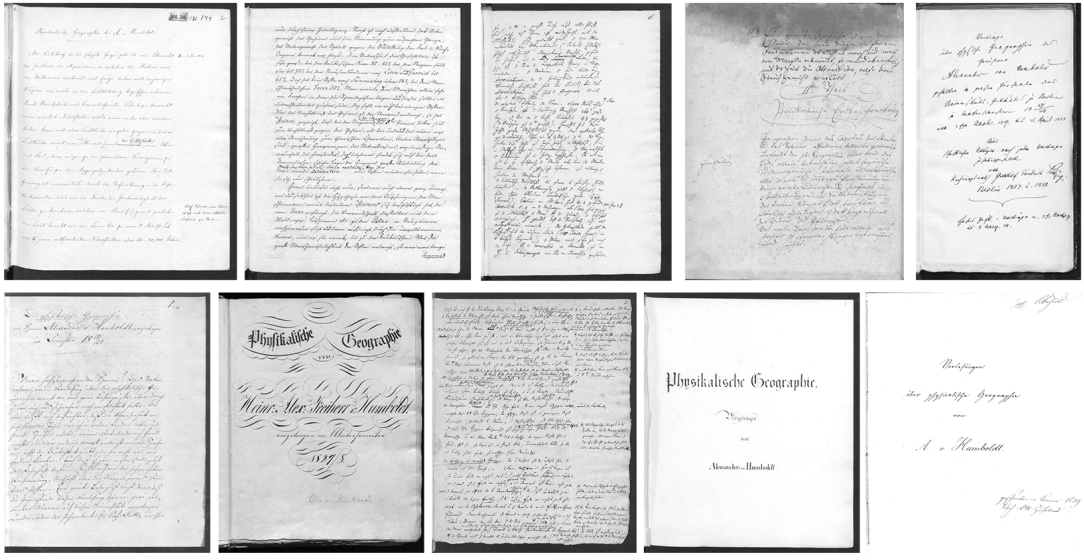


Figure 4: (Title) pages of ten currently known notebooks by Humboldt's listeners, from top left to bottom right, starting with the University lectures.[28]

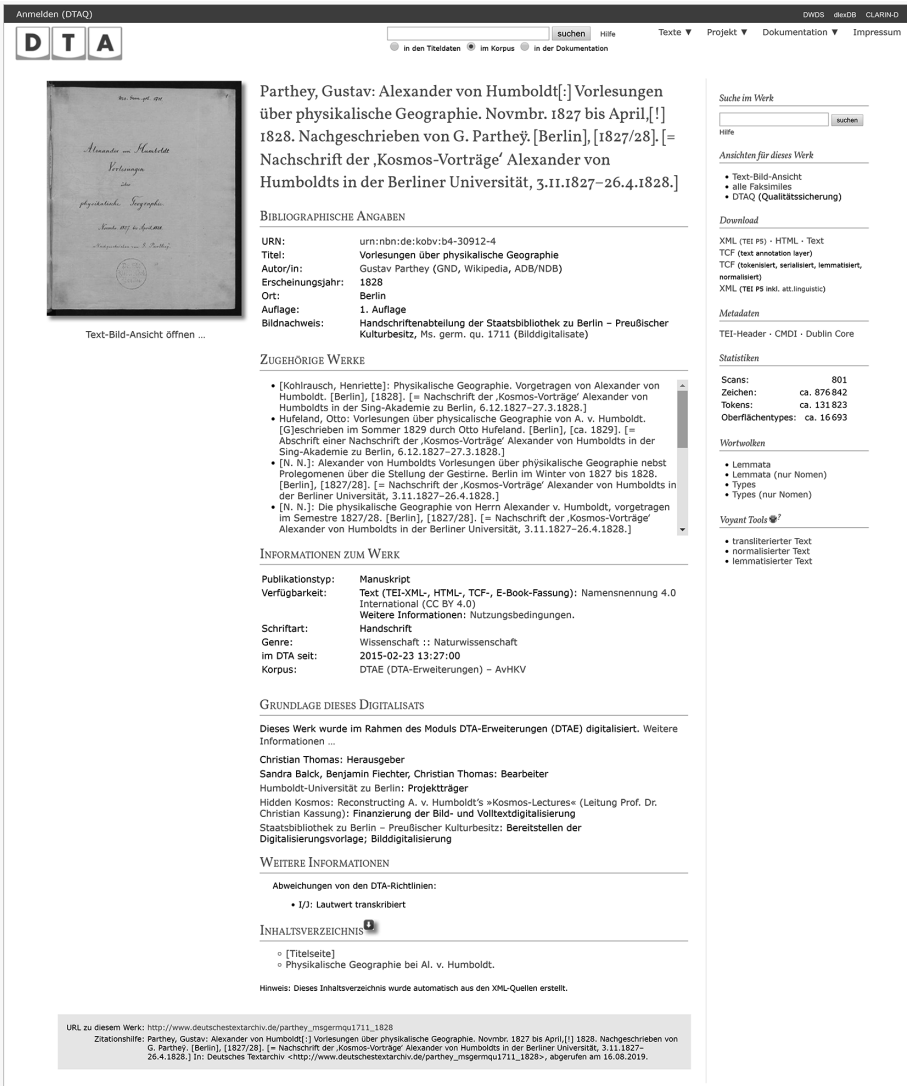


Figure 5: Screenshot of the *Deutsches Textarchiv* publication of one of the attendee's notebooks, Parthey.

these are the two notebooks that had been previously published in print editions, but for the online-edition the transcription has been collated against the handwritten original and corrected thoroughly. In total, the text corpus contains ca. 3,600 pages. The great individuality of these notebooks once again reminds us what it means to deal with a wealth of, in principle, equally valuable versions of an event. Their extent ranges from 80 (Riess) to 800 handwritten pages (Parthey; see figure 5).

The notebooks show a great individuality not only in this respect, but also concerning the narrative point of view (first person narrative vs. third person) and the degree to which the text was elaborated (from shorthand notes in keyword style to fully formulated sentences). Some documents were obviously produced in closer relation to each other, either by copying the whole notebook or by transposing single lessons or passages from one to the other. Other manuscripts are completely independent of each other and bear only little, but still palpable resemblance, like distant relatives of a big family. Some manuscripts were clearly edited by three or more different hands, some contain illustrations, some don't, and so on.

I argue that this individuality of each attendee's notebook is an inevitable effect of the circumstances of their production: The originators—who are not necessarily the scribes of the notebooks, since they may have hired a professional scribe to manufacture the fair copy—visited the lectures, took notes of the course and only some hours or even days later these notes were transformed into the running text of the fair copies handed down to us. This multistage process of mediation leads to the great diversity of the end products which is typical for this type of witnesses, and which must not (as would be tradition) be levelled out by constructing one definite (and definitely fictional) "ideal" text. Instead, each version has to be appreciated in its own right. This non-hierarchical parallelism of different versions of the Kosmos-Lectures has to be made visible and accessible, instead of being covered by the editor's interpolation of the (from his specific point of view) most likely account of the event. Important devices to reach this goal and to keep the multitude of witnesses manageable are document-spanning overviews of shared features such as the chronological order of the single lessons and the succeeding, but often implicit outline of topics dealt with over time, as well as accumulated, interactive lists of persons[31] or scientific instru-

subset of the TEI P5 tagset. See Haaf and Thomas as well as the documentation at www.deutschestextarchiv.de/doku/basisformat_manuskripte, for the annotation guidelines (work in progress). Accessed 2 Dec. 2018.

[31] Available at www.deutschestextarchiv.de/kosmos/person. Accessed 2 Dec. 2018. This overview contains all ca. 900 persons Humboldt mentioned during the courses, each entry related to an authorial database and linked to its context of appearance in the respective attendee's notebook. The list is generated directly from the TEI-XML-conformant encoding of occurrences of `<persName>`s within the source documents. This relatively simple database leads to observations and research questions comparing the two cycles of lectures with respect to the individual accounts of each cycle.

ments[32] mentioned in the different lessons. These document-spanning overviews offer means of orientation and facilitate the comparing, parallel study of the several witnesses.

Another method at hand is the computer-aided collation of notebooks (or sections thereof) that apparently have been copied from another manuscript, with their master copy. As it turns out, only one original notebook from the lectures at the Sing-Academy hall in 1827/28 is known at present ([N.N.] a), while the other manuscript covering this class ([N.N.] e; Hufeland) are merely copies of that original text. On the title page of the latter it is clearly stated that this manuscript was "geschrieben im Sommer 1829 durch Otto Hufeland" (Hufeland 3). While it would still be possible that this is an independent manuscript covering the lectures that was produced only in 1829, but based on notes taking during the course in 1827/28, the collation against [N.N.] a reveals it to be a copy (of a copy) of that manuscript[33]. Comparing these two notebooks with XML-aware collation software like *Juxta* or *CollateX*[34] reveals which passages were added by the copyist that are not contained in the master copy. These passages therefore do not represent parts of Humboldt's lectures, but embody later knowledge. Collating copy and master manuscript can also help to reveal misunderstandings and flaws in the original—and vice versa.

Another attendee, the anonymous scribe of [N.N.] c, seems to have missed a couple of lessons. In order to keep a complete record of the course, he apparently asked a fellow listener to help out with his notes. As a result, the notebook's text covering these (and only these) passages displays a much greater similarity than the rest of the witnesses when compared lesson by lesson. The lessons in question can be determined by relatively simple means: first the XML-annotated transcriptions of the whole notebooks are split into single lessons and then a robust and easy-to-use software like *WCopyfind*[35] is used to determine the degree of similarity among these

---

[32] Available at www.deutschestextarchiv.de/kosmos/instrument. Accessed 2 Dec. 2018. Each instrument is linked to the Wikipedia entry explaining its purpose, usage and history, as well as the source documents where Humboldt or his predecessors describe the instrument. The different synonyms or spelling variants of each instrument can be searched and lead directly to our central primary sources, i.e. the attendees' notebooks in the *Deutsches Textarchiv*. On this work in progress list, see Hug and Thomas.

[33] Only recently, in 2018, another notebook that had been sold at an auction in 2011 to an unknown person resurfaced when its current owner, Geir Stenmark, a private collector from Norway, got in touch with the author of this paper. Mr. Stenmark took it upon himself to digitise the complete notebook and generously agreed to the publication of the scans which now can be found on the BBAW's digilib-Server. This re-discovered notebook/copy, which is listed as [N.N.] e in the bibliographical section of this paper, helped to clear up the dependencies between the altogether three manuscripts covering the Sing-Academy lectures: It is now beyond doubt that [N.N.] a is the master script, of which [N.N.] e is a direct copy, while Hufeland is a copy of that copy.

[34] *Juxta Collation Software for Scholars*, www.juxtasoftware.org/; *CollateX – Software for Collating Textual Sources*, collatex.net/. Both accessed 2 Dec 2018. See Thomas for examples of the usage of these tools in the context of the *Hidden Kosmos*-project.

[35] *WCopyfind*, available via plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/. Accessed

documents. With some minor variance depending on the parameters used,[36] usually the documents display the great variety typical for this type of sources: only between 3 and 20% of the text of two documents compared is considered a match.

But for several distinct lessons the text from two manuscripts, [N.N.] c and Parthey, matches by up to 80 or even more than 90% (while other lessons between the same two manuscripts again show the typical low similarity of 3 to no more than 20%). This can only be explained by assuming that one scribe copied the lessons in question from the other, which is confirmed by a close reading comparison of the passages from both manuscripts. The fact that one source, Parthey, is always richer than the other, [N.N.] a, where the copyist left out certain passages to simplify and rush his task, shows that the latter has been copied from the former and not the other way around. Now we can continue to collate the two texts regarding the lessons we identified, while collating the remainder of the manuscripts would be pointless due to their high variance.

5 Humboldt as a DJ: Re-mixing the Kosmos-Lectures

By the same method (i.e. automated comparison of several witnesses), we can discover surprising similarities not only between the notebooks themselves, but also between these pivotal witnesses of the Kosmos-Lectures and other texts of Humboldt's. As is to be expected, the Kosmos-Lectures are a synopsis of Humboldt's own work and, of course, also that of his predecessors and contemporaries until 1827/28, when the lectures were given. We can infer from his legacy collection and, even more so, from the transcripts of his attendees that Humboldt used earlier works of his own and of his fellow scientists to prepare for the lectures. The notes of his auditors make it very likely that he even read out passages from previously printed works during some lessons. One of the most striking examples discovered in the context of the research presented here is Humboldt's lecture "Über die Hauptursachen der Temperatur-Verschiedenheit auf dem Erkörper," published in 1827 in Poggendorff's *Annalen der Physik*, and then again, in a slightly different version, in 1830 in the Prussian Academy's *Abhandlungen*. Humboldt presented this lecture to his colleagues at the Prussian Academy of Sciences in Berlin on July 3rd, 1827, i.e. only three months before the Kosmos-Lectures at the University started. A significant amount of text from this Academy lecture is echoed in several attendee's notebooks from the Kosmos-Lectures, resulting in an unusually high similarity between these notebooks pertaining to these passages. This can only be explained by the assumption that

2 Dec. 2018.

[36] I.e. if punctuation and upper case is ignored or not, numbers are included or excluded, the number of identical words that count as a match phrase is set greater or smaller, etc.

Humboldt simply read out the script he had elaborated for his Academy lecture in June some weeks later at the University, and again several more weeks later at his second course in the Sing-Academy building.

As I have already elaborated these findings elsewhere (Erdmann and Thomas), I will focus on their consequences for the conception of the Kosmos-Lectures as consisting of and surviving in multiple versions. The results indicate that Humboldt used significant parts of previous publications when preparing for the lectures, including his own published work as well as that of others and unpublished material that has been send to him by his fellow researchers. As a consequence, these documents have to be considered as vital parts of the multiple versions of the Kosmos-Lectures.

Neither were these lectures created *ex nihilo* nor was Humboldt simply done with the material base once the final lesson ended in April 1828: It is evident that he re-used his scripts as "raw material" for the *Kosmos* and other publications.[37] This has a considerable, yet usually under-appreciated effect on the "source' as well as on the "target" material: Humboldt (orally) re-published (parts of) previous texts as parts of singular lecture units, i.e. he integrated these parts into another, genuinely new and in itself complete publication: the Kosmos-Lectures. While doing so, he also changed the "register" or channel of communication from written text to speech, put the original text in a different context, and presented it to a different audience at a different time. Thereby, not only did he create another version of each of the original documents in question, but similar to popular remix culture, Humboldt created a whole new tune out of various samples lent from his own and other's previous works. Once the lectures had finished, he continued to sample the original Kosmos-Lectures into his following publications (see figure 6[38]).

---

[37] Exactly which previously published documents Humboldt used in preparation, and which ones he re-used afterwards for which publications (other than the *Kosmos*), to what extent, and with which alterations, still remains to be investigated.

[38] In chronological order, from top left to bottom right: *1)* Nachl. A. v. Humboldt, gr. K. 1, Mp. 2, Nr. 13, p. [1], resolver.staatsbibliothek-berlin.de/SBB000162A400000001: Table on (average) temperatures in Berlin by J. H. Mädler with notes by Humboldt, ca. 1825; *2) Temperatur-Verschiedenheit Annalen*, title page, gallica.bnf.fr/ark:/12148/bpt6k150967/f13.item: print of Humboldt's lecture "Über die Hauptursachen der Temperatur-Verschiedenheit auf dem Erkörper," presented at the Prussian Academy of Sciences in Berlin on July 3rd 1827 (See Thomas and Erdmann on this particular subject, i.e. Humboldt's ongoing occupation with the annual average temperatures and climate zones on the planet); *3)* Nachl. A. v. Humboldt, gr. Kasten 6, Nr. 13, Bl. 1r, resolver.staatsbibliothek-berlin.de/SBB00019EC800000000: Letter from K. A. Rudolphi to Humboldt, 7.11.1827, on the topic of intestinal worms (one of Rudolphi's favourite subjects); *4)* Nachl. A. v. Humboldt, gr. Kasten 11, Nr. 7, Bl. 3–15, Bl. 13r, resolver.staatsbibliothek-berlin.de/SBB0001AB8300000025: Note by Humboldt to himself [not dated, ca. 1827] to remind him to look up where Francis Bacon (much earlier than J. R. Forster) stated that all continents towards the south had a pyramidal shape; *5)* Patzig 2016, p. 291: one of the auditor's notebooks, 47th lesson, at which Humboldt clearly read out passages from item 2) and added some details he re-used in its re-publication as item 6); *6) Temperatur-Verschiedenheit Abhandlungen*: Title page of the revised re-publication of 2); *7)* Title page of *Kosmos. Entwurf*, I, in which Humboldt re-used and elaborated on 6). All URLs accessed

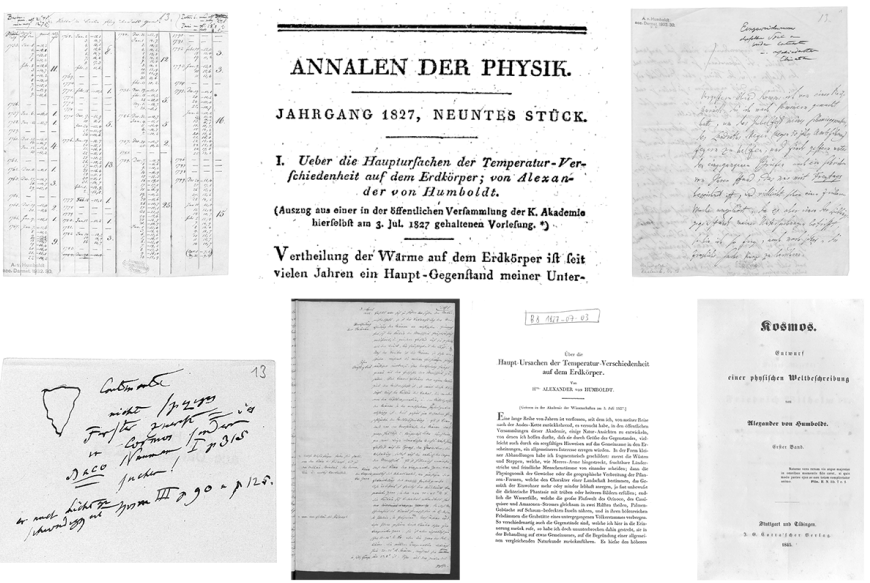


Figure 6: Some of the material Humboldt evidently re-used for/from his lectures.

They have become parts of a new whole, itself consisting of versions only and accessible to us only as such: as versions, competing with, contradicting and complementing each other. I believe that the theoretical reasoning and methodological implications derived from the exemplary corpus presented here are transferrable to many other, similar instances in which we are necessarily dealing with cultural history on the basis of various, but always "indirect" historical transmissions.

Bibliography

Primary sources related to Alexander von Humboldt's "Kosmos-Lectures"

N.N. a *Physikalische Geographie. Vorgetragen von Alexander von Humboldt. [Berlin], [1828]. [= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Sing-Akademie zu Berlin, 6.12.1827–27.3.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2015. www.deutschestextarchiv.de/nn_msgermqu2124_1827. Accessed 2 Dec. 2018.

N.N. b. *Die physikalische Geographie von Herrn Alexander v. Humboldt, vorgetragen im Semestre 1827/28. [Berlin], [1827/28]. [= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts*

2 Dec. 2018

in der Berliner Universität, 3.11.1827–26.4.1828.], edited by Christian Thomas, Deutsches Textarchiv, 2015. www.deutschestextarchiv.de/nn_oktavgfeo79_1828. Accessed 2 Dec. 2018.

N.N. c. *Alexander von Humboldts Vorlesungen über phÿsikalische Geographie nebst Prolegomenen über die Stellung der Gestirne. Berlin im Winter von 1827 bis 1828. [Berlin], [1827/28]. [= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2015. www.deutschestextarchiv.de/nn_msgermqu2345_1827. Accessed 2 Dec. 2018.

N.N. d. *Physikalische Geographie von Heinr. Alex. Freiherr v. Humboldt. [V]orgetragen im Wintersemester 1827/8. [Berlin], [1827/28]. [= Nach-schrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2015. www.deutschestextarchiv.de/dtaq/-book/view/nn_n0171w1_1828. Accessed 2 Dec. 2018.

N.N. e. *Phÿsikalische Geographie[.] Vorgetragen von Alexander von Humboldt. Angefangen d[en] 6ten Xbre 1827. Abschrift des Heftes der Frau Geheimräthin Kohlrausch. [= Abschrift einer Nachschrift der ‚Kosmos-Vorträge Alexander von Humboldts in der Sing-Akademie zu Berlin, 6.12.1827–27.3.1828.]* [Berlin], [ca. 1828]. digilib-Server der Berlin-Brandenburgischen Akademie der Wissenschaften, 2018, digilib.bbaw.de/digitallibrary/jquery/digicat.html?fn=/silo10/avhr/Humboldt_Kosmos-Vortraege-SingAkademie_1828-Abschrift_2018-11-12. Accessed 27 Dec. 2018.

Hufeland, Otto. *Vorlesungen über physicalische Geographie von A. v. Humboldt. [G]eschrieben im Sommer 1829 durch Otto Hufeland. [Berlin], [ca. 1829]. [= Abschrift einer Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Sing-Akademie zu Berlin, 6.12.1827–27.3.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2015. http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829. Accessed 2 Dec. 2018.

Libelt, Karol. *Wykłady Humboldta na uniwersytecie Berlińskim: notaty prelekcyj tych po uczniu Jego Karolu Libelcie. [s. l.], [1828]. [= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2015. www.deutschestextarchiv.de/libelt_hs6623ii_1828. Accessed 2 Dec. 2018.

Lohde, Ludwig. *Physikalische Geographie. Eine Vorlesung des Herrn A. v. Humboldt[,] gehalten im Winter 1827.* Digitalisierte Sammlungen der Staatsbibliothek zu Berlin, 2015. resolver.staatsbibliothek-berlin.de/SBB00016BAE00000000. Accessed 2 Dec. 2018.

Parthey, Gustav. *Alexander von Humboldt[:] Vorlesungen über physikalische Geographie. Novmbr. 1827 bis April,[!] 1828. Nachgeschrieben von G. Partheÿ. [Berlin], [1827/28]. [= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2015. www.deutschestextarchiv.de/-parthey_msgermqu1711_1828. Accessed 2 Dec. 2018.

Patzig, Gotthilf. *Vorträge über physische Geographie des Freiherrn Alexander von Humbold[!]: gehalten im großen Hörsaale des Universitäts-Gebäudes zu Berlin im Wintersemester 1827/28 vom 3ten Novbr. 1827. bis 26 April 1828. Aus schriftlichen Notizen nach jedem Vortrage zusammengestellt vom Rechnungsrath Gotthilf Friedrich Patzig. Berlin, 1827/28. [= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner*

Universität, 3.11.1827–26.4.1828], edited by Christian Thomas, Deutsches Textarchiv, 2016, www.deutschestextarchiv.de/patzig_msgermfol841842_1828. Accessed 2 Dec. 2018.

Riess, Peter Theophil. *Physikalische Geographie nach Al. v. Humboldt. [Berlin], [1827/28]. [= Nachschrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2016. www.deutschestextarchiv.de/riess_f2e1853_1828. Accessed 2 Dec. 2018.

Willisen, Friedrich Adolf von. *Humbolds[!] Vorlesungen. [Berlin], [1827/28]. [= Nach-schrift der ‚Kosmos-Vorträge' Alexander von Humboldts in der Berliner Universität, 3.11.1827–26.4.1828.]*, edited by Christian Thomas, Deutsches Textarchiv, 2017. www.deutschestextarchiv.de/willisen_humboldt_1827. Accessed 2 Dec. 2018.

Secondary literature

Alexander von Humboldts Vorlesungen über physikalische Geographie nebst Prolegomenen über die Stellung der Gestirne. Berlin im Winter von 1827 bis 1828. Erstmalige (unveränderte) Veröffentlichung einer im Besitze des Verlages befindlichen Kollegnachschrift. Miron Goldstein, 1934.

Alexander von Humboldt. Eine wissenschaftliche Biographie, edited by Karl Bruhns. Brockhaus, 1872. 3 vols.

Briefe von Alexander von Humboldt an Varnhagen von Ense, aus den Jahren 1827 bis 1858. Nebst Auszügen aus Varnhagen's Tagebüchern, und Briefen von Varnhagen und Andern an Humboldt, edited by Ludmilla Assing, Brockhaus, 1860.

Die Vorlesungen der Berliner Universität 1810–1834 nach dem deutschen und lateinischen Lektionskatalog sowie den Ministerialakten, edited by Wolfgang Virmond, Akademie Verlag, 2011.

Engelmann, Gerhard. *Die Hochschulgeographie in Preußen 1810–1914.* (*Erdkundliches Wissen*, H. 64.). Franz Steiner Verlag, 1983.

Erdmann, Dominik, and Christian Thomas. "„… zu den wunderlichsten Schlangen der Gelehrsamkeit zusammengegliedert". Neue Materialien zu den ‚Kosmos-Vorträgent' Alexander von Humboldts, nebst Vorüberlegungen zu deren digitaler Edition." *HiN – Humboldt im Netz. Internationale Zeitschrift für Humboldt-Studien*, XV, 28, 2014, pp. 34–45. DOI: 10.18443/189. Accessed 2 Dec. 2018.

Erdmann, Dominik, and Jutta Weber. "Nachlassgeschichten – Bemerkungen zu Humboldts nachgelassenen Papieren in der Berliner Staatsbibliothek und der Biblioteka Jagiellońska Krakau." *HiN – Humboldt im Netz. Internationale Zeitschrift für Humboldt-Studien*, XVI, 31, 2015, pp. 58–77. DOI: 10.18443/223. Accessed 2 Dec. 2018.

Fischer, Christian August. *Ueber Collegien und Collegienhefte: oder: erprobte Anleitung zum zweckmäßigsten Hören und Nachschreiben sowohl der Academischen als der höheren Gymnasial-Vorlesungen.* Habicht, 1826.

Gabriele von Bülow – Tochter Wilhelm von Humboldts. Ein Lebensbild aus den Familienpapieren Wilhelm von Humboldts und seiner Kinder 1791–1887, edited by Anna von Sydow, Mittler und Sohn, 1893.

Haaf, Susanne, and Christian Thomas. "Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format." *Journal of the Text Encoding Initiative*, Issue

10, 2016, online since 08 August 2017. DOI: 10.4000/jtei.1650. Accessed 2 Dec. 2018.

Hug, Marius, and Christian Thomas. "Den Kosmos sondieren: Das Thermometer und andere Instrumente der Wissenschafts- und Technikgeschichte in A. v. Humboldts ‚Kosmos-Vorträgen'." Presentation at the Workshop *Wissenschaftsgeschichte und Digital Humanities in Forschung und Lehre*, 07.04. bis 09.04.2016 in Göttingen, 2016. Humboldt University of Berlin, www.culture.hu-berlin.de/de/forschung/projekte/hidden-kosmos/media/avhkv-thermometer.pdf. Accessed 2 Dec. 2018.

Humboldt, Alexander von. "Über die Hauptursachen der Temperatur-Verschiedenheit auf dem Erdkörper." *Annalen der Physik,* edited by J. C. Poggendorff, 11, 1827, pp. [1]–27.

———. "Über die Haupt-Ursachen der Temperatur-Verschiedenheit auf dem Erd-körper." *Abhandlungen der Königlich Preussischen Akademie der Wissenschaften in Berlin. Aus dem Jahre 1827.* F. Dümmler, 1830, pp. 295–316.

———. *Kosmos. Entwurf einer physischen Weltbeschreibung.* 5 vol. Cotta, 1845–62.

———. *Cosmos. A Sketch of the Physical Description of the Universe.* Translated from the German, by E. C. Otté. Volume 1. Bohn, 1848.

———. *Über das Universum. Die Kosmosvorträge 1827/28 in der Berliner Singakademie,* edited by Jürgen Hamel and Klaus-Harro Tiemann. Insel, 1993.

———, and Henriette Kohlrausch. *Die Kosmos-Vorlesung an der Berliner Sing-Akademie*, edited by Christian Kassung and Christian Thomas. Insel, 2019.

Klein, Ursula. *Humboldts Preußen. Wissenschaft und Technik im Aufbruch.* Wissenschaftliche Buchgesellschaft, 2015.

Thomas, Christian. "Hidden Kosmos – Humboldts ‚Kosmos-Vorträge' als Probe der Digital Humanities." *Book of Abstracts (Vorträge) zur DHd-Jahrestagung 2015*, gams.uni-graz.at/o:dhd2015.abstracts-vortraege, pp. 193–201. Accessed 2 Dec. 2018.

Thomas, Christian, and Dominik Erdmann. "From the Concept of Isothermal Lines (1817) to the Principal Causes of the Difference of Temperature on the Globe (1827–1830–1855) – A Curated Reading Example." Berlin, 2015. Available at Humboldt-Universität zu Berlin, www.culture.hu-berlin.de/de/forschung/projekte/hidden-kosmos/media/thomas-erdmann-2015-reading-example.pdf. Accessed 2 Dec. 2018.

Thomas, Christian, et al. "Methoden und Ziele der Erschließung handschriftlicher Quellen zu Alexander von Humboldts Kosmos-Vorträgen: Das Projekt Hidden Kosmos der Humboldt-Universität zu Berlin." *Horizonte der Humboldtforschung,* edited by Ottmar Ette and Julian Drews. Olms-Weidmann, 2016, pp. 287–318.

Verzeichniß der Vorlesungen, welche von der Universität zu Berlin im Winterhalbenjahre 1827 bis 1828 vom 22. Oktober an gehalten werden. [1827].

Werner, Petra. *Himmel und Erde: Alexander von Humboldt und sein Kosmos.* (*Beiträge zur Alexander-von-Humboldt-Forschung*, 24.) Akademie Verlag, 2004.

Wulf, Andrea. *The Invention of Nature: Alexander von Humboldt's New World.* Knopf, 2015.

Ontological Approaches to Versioning

Versioning Materiality: Documenting Evidence of Past Binding Structures

Athanasios Velios and Nicholas Pickwoad

Abstract

Describing the structure and materials of bookbindings is an essential task of the study of the history of the book. Books with repaired or replaced binding structures are of particular interest, given that evidence of one or even two or more previous structures often remain on the book. The results of rebinding can be considered as separate versions of the binding structure. Evidence of the binding structures need to be matched with the corresponding version of the binding. This helps in formulating provenance.

In this paper we discuss problems of documenting binding evidence including a) the reuse of earlier components in later bindings and b) the reuse of components originally belonging to other books. After a review of different approaches to the description of earlier bindings we focus on the CIDOC CRM as a possible way of modelling the versions of bindings through an event-centric approach and offering a number of examples. Finally, we discuss the advantages of using the CRM for versioning as well as the limitations of our method.

1 Introduction

By *versioning* we often mean keeping track of the changes of text (e.g. different versions of a report). In computer programming, a plethora of tools allow changes in programming code to be tracked. Versioning allows developers to follow the history of a file over long periods.

In other fields of research, tracking the changes of material objects during their history is common practice. In archaeology, art history, conservation and other relevant fields, understanding changes to material objects leads to conclusions about their technology and use. In this paper, we propose the adoption of the idea of versioning to the description of material objects in order to capture the changing nature of an object over the centuries.

This is particularly important in the case of historic books. The book as a material object is a representation of the social, economic, and cultural environment in which it was produced or modified (McKenzie; Darnton), because it can combine a variety of crafts (including sewing, carpentry, leatherwork, embroidery, and gilding) and a variety of materials (from parchment to metal).

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 103–126.

1.1 CIDOC CRM

The Conceptual Reference Model (CRM) published by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) has been an important influence in the development of this work. The CIDOC CRM (ISO 21127:2006) is a formal ontology. It defines concepts (*entities*) and relationships (*properties*) within the cultural heritage sector to model relevant activities. These entities are organised in hierarchies from the more general to the more specific. Generic entities (*parent entities*) contain more specific entities (*child entities*). Any child entity shares the characteristics of its more general parent entity. For example the entity *E5 Event* is the parent of both *E67 Birth* and *E69 Death*. *E67 Birth* is an *E5 Event*, but clearly not all *E5 Event*s are *E67 Birth*s, since we also have *E69 Death* among other types. The hierarchy formed with parent and child entities is often called an *IsA* hierarchy. Also, any characteristics of *E5 Event* (e.g. the fact that people participate in events) are also applicable to the child entities. This is also known as *property inheritance*. For an introduction to the CIDOC CRM, see Doerr ("The CIDOC Conceptual Reference Model").

The CIDOC CRM has been tested successfully for many years, resulting in a stable model. Because of this stability, the CIDOC CRM could be used as an abstract blueprint structure for documentation systems. We have adopted it here to demonstrate our use of versioning.

The paper begins with some background information about historic bookbinding and the documentation of binding structures. It then introduces concepts from the CIDOC CRM which are relevant to versioning bindings and it proposes a structure that can be adopted to document them. It examines a case study demonstrating the principles of that structure and it concludes with some points for discussion. Some bookbinding terms used in this paper may be unfamiliar to the reader. We are using these terms in italics followed by a citation to the term in brackets and single quotes. These are included in the references.

2 Dating bindings

Bookbindings are frequently ignored in descriptions of books and in library catalogue entries. However they often carry important information about where books have been, and therefore where they may have been read. This can be done by establishing chronological and geographical ranges for the use of particular techniques, materials and styles of decoration. The *textblocks* (LoB, "Textblocks") of books frequently have a longer life than their bindings. Their bindings are often either repaired or (partly) replaced. The ability, therefore to identify and date these sequences of binding,

rebinding and repair is critical to our understanding of the histories of individual books.

2.1 Rebinding books

Books would typically be rebound or repaired in response to damage or changes of fashion. For example, *covers* (LoB, "Covers") would be replaced when a library or a collector decided to update the appearance of their books. This would be done by completely replacing existing bindings, in which case only the *sewing stations* (LoB, "Sewing Stations") of the original structure will survive, or by replacing or covering the original covers with a new, perhaps more fashionable material, in which case the binding may well retain a first structure under a later covering. This has happened in many libraries, such as the collection of very early manuscripts in the Biblioteca Capitolare in Vercelli (Lombardia), where the original full covers were mostly replaced by *quarter covers* (LoB, "Quarter Covers") of *tanned* (LoB, "Tanned Skin") *sheepskin* (LoB, "Sheepskin") in the late seventeenth century, or the library of the Franciscan monastery of Šibenik in Croatia, where both the *boards* (LoB, "Boards") and covers of the bindings of their collection of incunabula were replaced in the eighteenth century by *laced-case* (LoB, "Laced Cases") covers of thick *cartonnage* (LoB, "Cartonnage") paper.

2.2 Reuse of components

During rebinding or repair, binders often used recycled material, mostly from earlier books. For example, printed or written leaves from earlier books recycled as *endleaves* (LoB, "Endleaves"), covers, *spine linings* (LoB, "Spine Linings"), *board laminates* (LoB, "Board Laminates"), etc., or boards and covers from discarded or earlier bindings which were recycled for different books. Any description of bindings that attempts to date them based on these materials may therefore be misleading, as there may be a discrepancy of several centuries between the materials used. As the result of this phenomenon, a Romanesque manuscript in the library of Lincoln Cathedral now has two wooden boards of the same age as the manuscript, neither of which matches either the manuscript or each other, but both of which were used in the repair of the book in the nineteenth century.

2.3 Case study

A copy of Jacobus Philippus, *De claris mulieribus*, Ferrara, 1497 (figure 1), once in the Otto Schäfer collection in Germany, was described in an exhibition catalogue (Arnim) as having been bound in a contemporary binding with a cartonnage cover attached by lacing the *slips* (LoB, "Slips") of the leather sewing supports through its *joints*

(LoB, "Joints (Features)"). This type of Italian laced-case cover is frequently found on bindings from the second half of the sixteenth century through the nineteenth century and if this binding were of the date of the text then it would be the earliest example known by almost half a century. A response from the author of the catalogue confirmed that the slips were part of the sewing supports and were original to the binding. An examination of the book in person a few years later in New York led to these observations:

a) the existence of leather stains at the *head* (LoB, "Head") and *tail* (LoB, "Tail") of the spine edges of the outermost endleaf at each end of the book,
b) the cut ends of substantial white *alum-tawed* (LoB, "Alum-Tawed Skin") *split-strap* sewing supports (LoB, "Split-Strap Sewing Supports") showing in the joints and
c) the existence of a multiplicity of worm holes in the first and last few leaves.

These observations indicate that the book was first bound in a contemporary *inboard binding* (LoB, "Inboard Bindings") with beech-wood boards (hence the wormholes—woodworms love beech wood) and a quarter cover of a dark reddish-brown tanned goatskin (hence the leather stains at the spine edge of the endleaves) of a typically Italian type (e.g. figure 2). The slips of tanned skin laced though the paper cover were in fact laced under the original alum-tawed sewing supports circa 1600 to attach the new cover, possibly to replace the earlier worm-damaged boards. A drawing with this evidence is shown in figure 3. Because this sequence of events was not first identified and recorded, the binding was inaccurately described and its description was misleading. In section 3.3 we explain how a data structure based on the idea of versioning can be used to capture the multiple components from different periods on this book. We first introduce non-structured documentation records to show how traditional methods of record keeping are inadequate.

3 Records of bindings

3.1 Free-text records

As mentioned in the example of the Arnim catalogue, bookbinding descriptions are often produced using free text (i.e. in prose). This is because free text has been well-rooted as a documentation tool in relevant fields such as palaeography and conservation (approaches such as this by Campagnolo or Stokes et al. who employ structured records are still exceptions in the respective fields). Free text offers an immediate narrative which can be easily followed by a reader. It inherits the flexibility of spoken language and therefore it can be tailored to different audiences. A condition report of a binding, written by a conservator for other conservators, will be very different to an auction catalogue description written by an auctioneer for possible

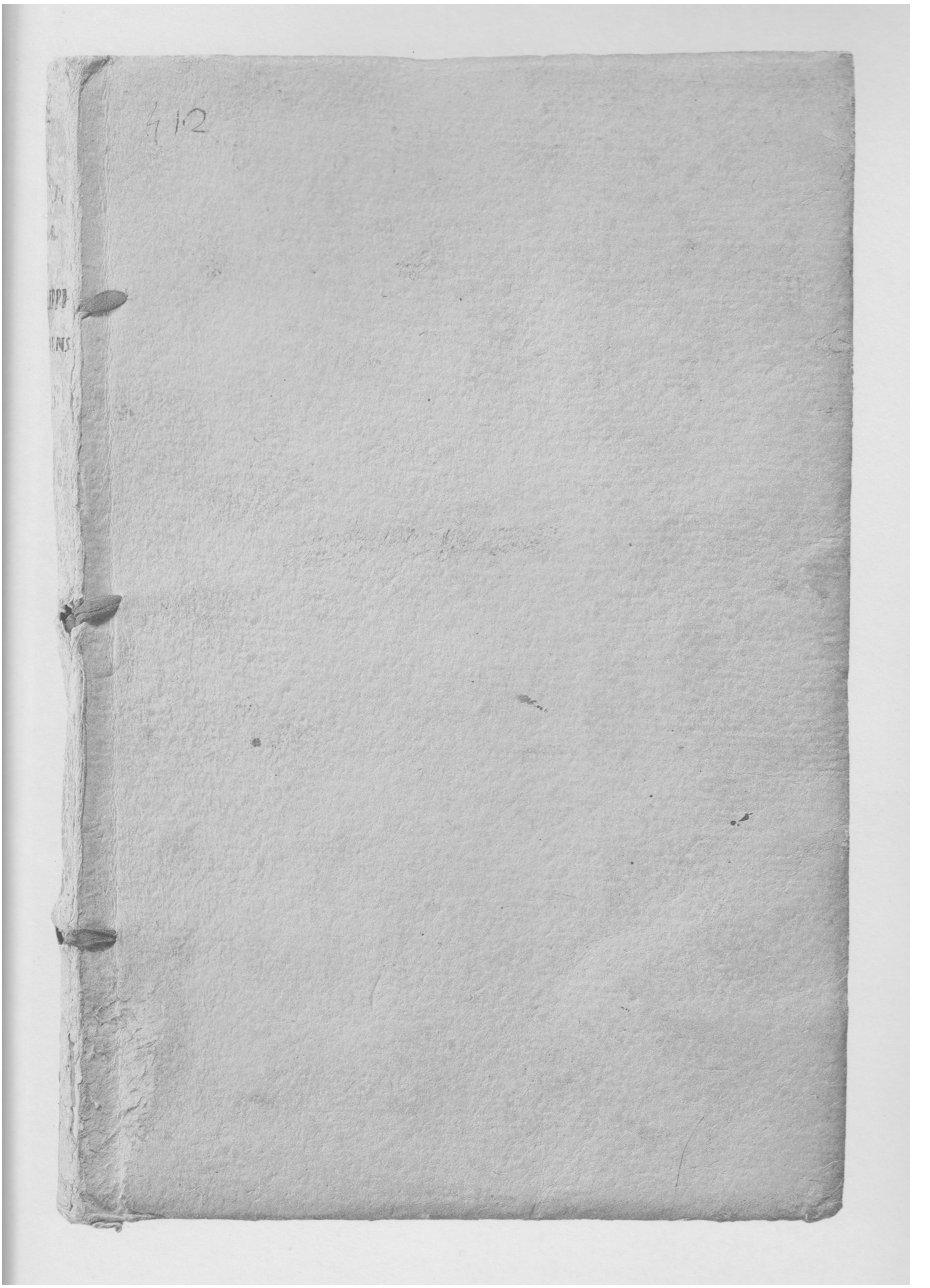


Figure 1: Photo of Jacobus Philippus, *De claris mulieribus*, Ferrara, 1497.

Figure 2: A typical Italian binding with a quarter cover over wooden boards from the late 15th century on a copy of: Ioannes a Sancto Geminiano, *Summa de exemplis*, Venice, 1499 (by permission of the Biblioteca di San Francesco della Vigna, Castello, Venice, RARI-B.III.13).

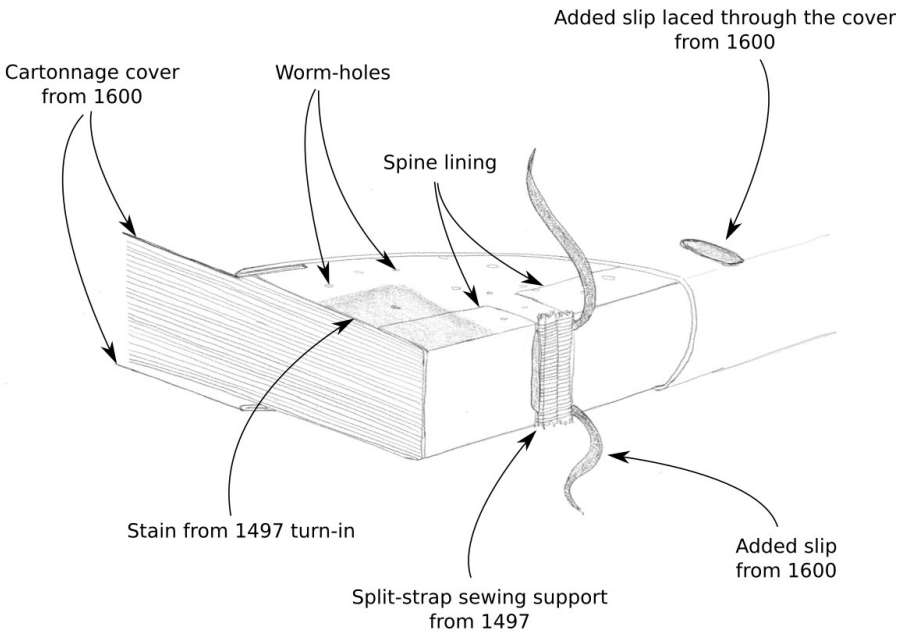


Figure 3: Drawing of evidence from a copy of Jacobus Philippus, *De claris mulieribus*, Ferrara, 1497.

collectors/bidders. The free-text description of the changes on a book can tell the history of the specific book. To build a picture about a collection or a period, a researcher needs to interpret free text descriptions and insert important observations in a database to improve the capacity for searching and summarising data. This interpretation leads to structured data.

In the field of historic bookbinding, descriptions of bindings with structured data require typologies, i.e. lists of terms corresponding to varying characteristics of bookbindings as we explain next.

3.2 Structured records – types

A number of projects and researchers have adopted structured records for bookbinding descriptions because they allow easier summary of data. These are typically in the form of a question being represented by a field, to which an answer can be given from a list of options. For example, the field *left board material* corresponds to the question "what material is the left board made of?" and the possible entries/answers

can be *wood*, *paper*, *tanned skin*, *alum-tawed skin*, etc. These terms define the types of material that a board can be made of. Ideally they should be organised as lists of terms in a controlled vocabulary or thesaurus where they can be retrieved through a lookup mechanism. When researchers retrieve types from the same controlled vocabulary or thesaurus, then it is possible to cross-search records from different collections. Examples of such vocabularies and thesauri are the thesaurus of the Rare Books and Manuscript Section of the Association of College and Research Libraries of the American Library Association (RBMS) and more recently the Language of Bindings Thesaurus (LoB).

The choice of fields/questions included in a structured record depends on its extent. Some records include hundreds of fields, such as the Saint Catherine Library survey (Velios and Pickwoad), while others include a small number of particularly significant fields such as the Wellcome Trust digitisation survey (Boal et al.). Most of these records focus on the current state of the binding, i.e. they include terms which describe the structure of the binding as it is at the time of the survey and not at the time that the binding was made. For example, it is expected to describe non-original *secondary covers* (LoB, "Secondary Covers") even if a binding only had a *primary cover* (LoB, "Primary Covers") when it was put together. This is useful for an accurate picture of the history of the object and for assessing the value of each binding component. The terms *primary* and *secondary cover* denote different types of covers based on the time that the cover was attached to the binding and define types of components based on time attributes, i.e. original or added at a later stage. Other examples are a) the distinct type of endleaves, called *inserted endleaves* (LoB, "Inserted Endleaves") which are defined as those which were added at a later stage, and b) the type of sewing for books that have been sewn more than once, which can be described as *current*, *previous* or *early*, depending on when each sewing was applied. There are two limitations when using types to describe time-related attributes of components:

1. Binding components added at different stages are mistakenly grouped together. In the example of the *inserted endleaves*, we may have two or more sets of endleaves added to a book at different times following the original binding. If we call all of them *inserted endleaves* we have no way to distinguish which set was first and which set followed.
2. Terms are arbitrarily created to cover earlier changes to a binding. In the example of the sewing structure we have allowed for the book to be bound up to three times (1 current, 2 previous, 3 earlier). How can we then describe the rare occasion where an even earlier fourth set of holes exists?

In the next section we will show a model for data structures which includes the sequence of events as opposed to implying it in types.

Previous experience

Binding survey work requires both direct observation of the current state of the binding and deductive thinking based on previous experience and understanding of binding structures. An experienced researcher is able to characterise evidence of absent components because of previous observation of such components on other bindings. To follow an earlier example, a set of currently unused sewing holes on the textblock is a strong indication that the book was bound using those holes in the past and that later it was rebound with the current set. The impression of a now missing thread in the spine fold of a bifolium between two unused holes is evidence of a thread once being present. Although the earlier sewing is not there, it is still possible to create a record of it through deduction. Therefore deduction is already an important process when creating structured records of bindings and often it is interlinked with observation. We will return to this issue in the next section and also in section 5.2.

The definitions of types of components include concepts of time and sequence of events. The use of such terms requires both the observation of remaining evidence from a removed component and the deduction of the type of that component based on previous observations. In the next section we propose a way to formalise the expression of time in bookbinding description using the CIDOC CRM.

3.3 Event-based records

In the previous section we explained that although the intention of bookbinding surveys may be to produce records of the state of the bindings at the time of the survey, they are also used to produce historical records of earlier states of the binding, through observation and deduction. We explained the limitations of object-centric records associated with terms. There is an important shift in the way that records of bindings should be conceived with the aim of overcoming these limitations: we are observing objects and deducing events that happened to these objects and therefore we should be creating records of events alongside records of objects. Events and objects are linked. Any event which may concern the history of a binding involves the object itself. The concept of a binding is persistent during the centuries of its history – it is the same object now as the one that the bookbinder created despite the many changes of its structure.

This leads to the question: when is a binding produced? Which events led to the production of the binding as a persistent object that we recognise and identify today and which events are modifications of that object? In many cases bindings were produced in stages. For example: often, a textblock would receive a temporary

stitched binding (LoB, "Stitched Bindings") soon after printing. At a later stage it would have been bound with a more permanent binding at the order of a customer. It is likely that a researcher will consider the event of adding the permanent binding to the textblock as the point where the binding for this object was produced. Another researcher may be particularly interested in temporary bindings and therefore would consider the stitched binding as the point in time when the binding was produced. We could consider the point of the production of the binding as a subjective choice of the researcher but in general it is safer to consider the earliest evidence of an action involving the textblock with the intention to keep the leaves together as the point where the object is produced. This means that from that point onward an identifier can be assigned to the object which can be used for reference.

Word lists and vocabularies used in the domain tend to focus more on the types of persistent items, i.e. the binding and its components and less on events and actions which are necessary to describe what happened to the object. The concept of the technique describes the making of an object, but in bookbinding descriptions it is considered as a characteristic of the object (and not of the making of the object). The LoB thesaurus includes hierarchies for both types of components and types of techniques. The intention of the thesaurus is that techniques should not be used to describe persistent items (bindings) but instead temporal items (events). The LoB thesaurus has been built based on the philosophy of the CIDOC CRM which is event-centric and a good candidate for describing the historical development of a binding.

After the production of a binding, there is a continuous timeline which we can use to describe the events that make up its history. Our observations reveal evidence from some of these events (a subset): those with the strongest impact are the *critical events*. In the same way that we may consider the starting point of the timeline subjectively, we may also consider the critical events subjectively based on previous experience. The records corresponding to the state of the object after each critical event can be considered as different versions of the binding.

Figure 4 shows an example of how CIDOC CRM entities can be used to build a timeline for a binding. Further references to other entities will be made later in this document. Temporal entities describe events of the book while persistent entities describe physical components. The thick arrows indicate an IsA hierarchy. The properties of each entity are shown linking two entities with a normal arrow. Properties of the higher entities are inherited by the lower entities.

The starting point of the history of a binding can be considered as an *E12 Production* which links with *E24 Physical Man-Made Thing* (the binding) through property *P108 has produced*. At the same time *E12 Production* is an *E11 Modification* and therefore inherits the property *P31 has modified* which can be used to describe the fact that components (*E24 Physical Man-Made Thing*s) were formed in advance of the binding

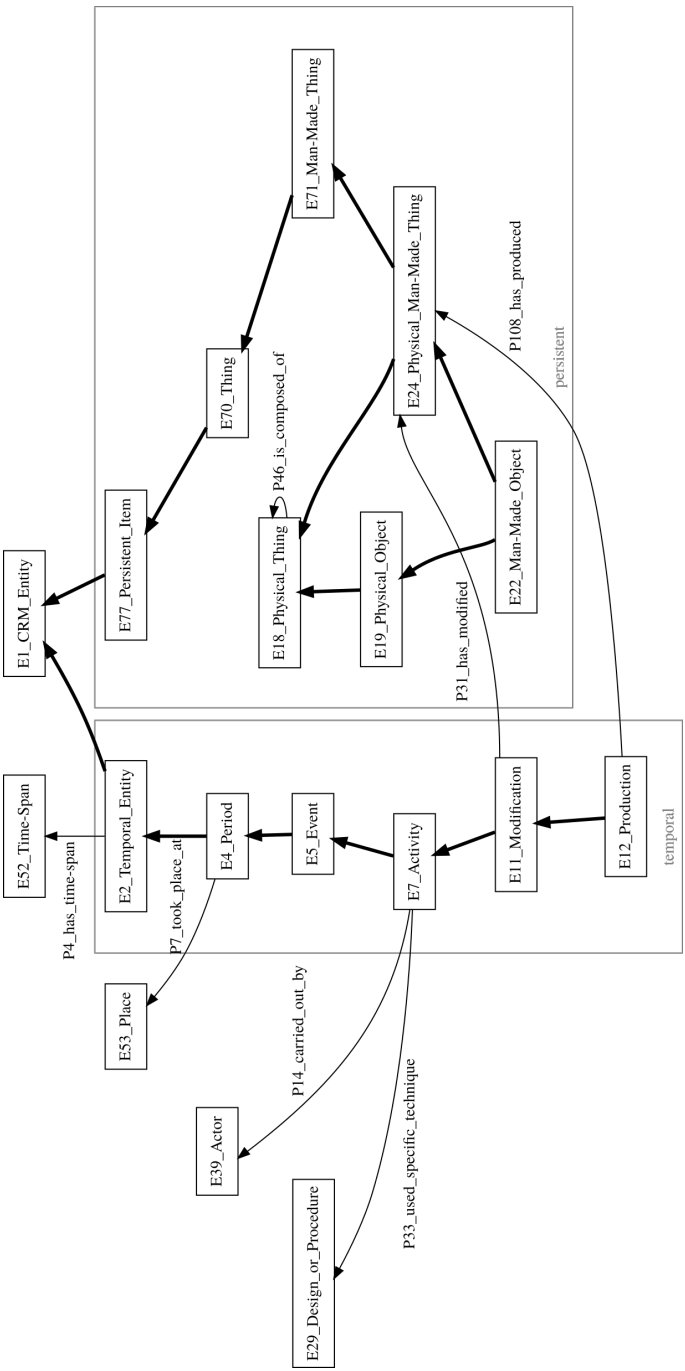


Figure 4: Selection of CIDOC CRM entities and properties for bookbinding timelines.

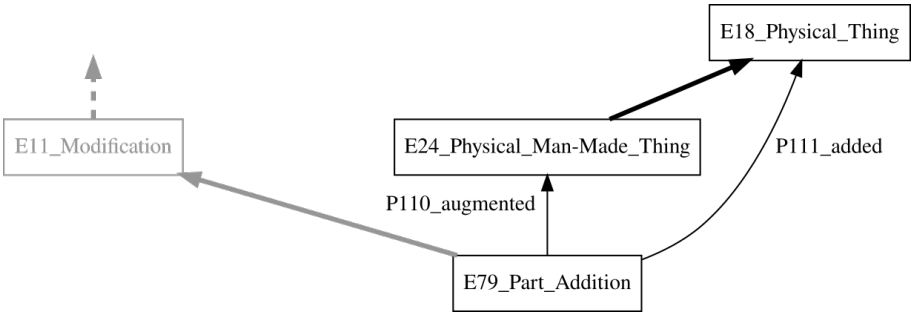


Figure 5: Selection of CIDOC CRM entities and properties for adding components to bindings.

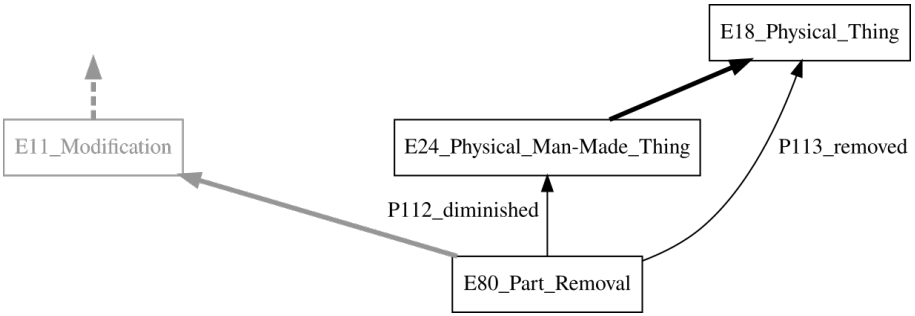


Figure 6: Selection of CIDOC CRM entities and properties for removing components to bindings.

of the book and were then used during the binding process. Higher up the temporal group of entities, we can use the properties: a) *P14 carried out by* to indicate the person or workshop that undertook the binding, b) *P33 used specific technique* to indicate the type of the technique used, c) *P7 took place at* to indicate where the creation of the binding happened and d) *P4 has time-span* to indicate the period that we have established as time that the binding was put together.

Further modifications to the binding at the various critical events can be modelled as shown in figure 5. To make the figure more legible, we have removed the groupings and the parent entities in the persistent entities group. *E79 Part Addition,* which is a modification, features two properties: a) *P110 augmented*, indicating the binding which was altered because of an addition of a new component and b) *P111 added*, indicating the component which was added (e.g. a new set of endleaves). All properties from the higher entities still apply so we can mark this modification as an event at a different time-span and by a different bookbinder or workshop.

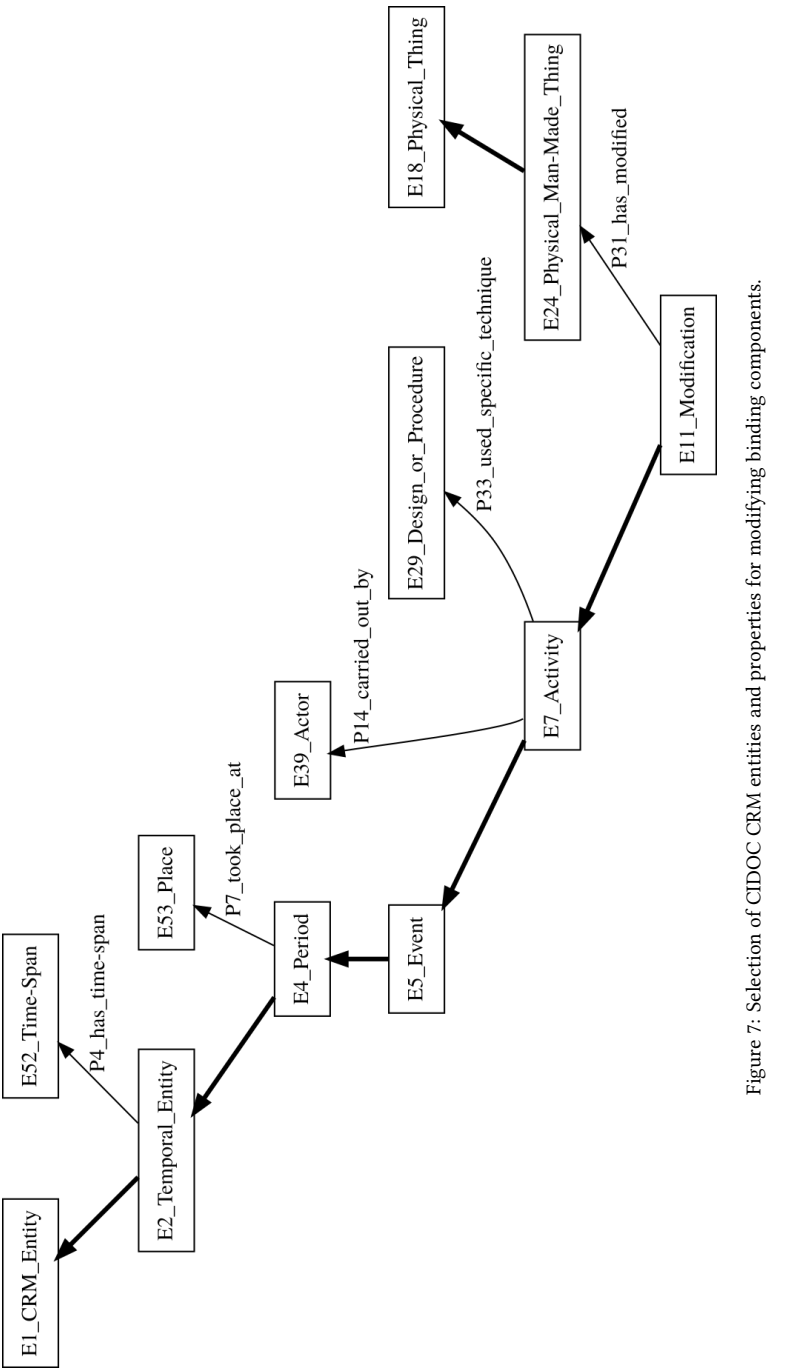


Figure 7: Selection of CIDOC CRM entities and properties for modifying binding components.

Figure 6 shows a similar arrangement of properties for removal (*E80 Part Removal*) of components from the binding, using the properties *P112 diminished* and *P113 removed* (e.g. the removal of a cover prior to it being replaced by another). Figure 7 shows a more generic structure for modifications of the binding which cannot be considered as either additions or removals.

Previous research by Ravenberg has shown that any change in a binding structure during conservation can be modelled by one of three options: an addition, a removal or a modification. We can also apply the same principle to any historic modification of the binding and therefore by modelling these three options we can arguably cover most of the historical activity affecting an object.

Each of these modification events can be considered to mark different versions of the binding. These events can be assigned an identifier and therefore references to the corresponding versions are then possible. In the next section we demonstrate the kind of records which can be produced for the various versions of the bindings of the case study book.

4 Case study

In the example we described in section 2.3, observed evidence indicates that the book has had two critical events during its history: the first binding around 1497 and the later covering around 1600. Figure 8 shows a basic CIDOC CRM structure we could use to map these events while recognising that there are other equally valid structures. The two binding events are at the bottom of the figure occupying different time-spans but both linked to our case study book. The property *P46 is composed of* is used to relate the book to its individual components. At this stage we make no statements about the period during which each component was present on the book. Much of the description of the book and components is done using terms from the LoB thesaurus and the property *P2 has type*. Even though the book no longer has its original boards, it can still be described as an *inboard binding* because the evidence is there to prove that the type *inboard binding* is applicable despite the fact that the boards are now missing. In the next section we will discuss the detailed expression of the activities altering the main components of the book and assigning periods to the existence of each component.

4.1 Boards

We consider an *E79 Part Addition* event labelled as *V1 Board addition*. The property *P117 occurs during* expresses the fact that the board was added while the event of binding was taking place. The property linking the event of adding the boards to the

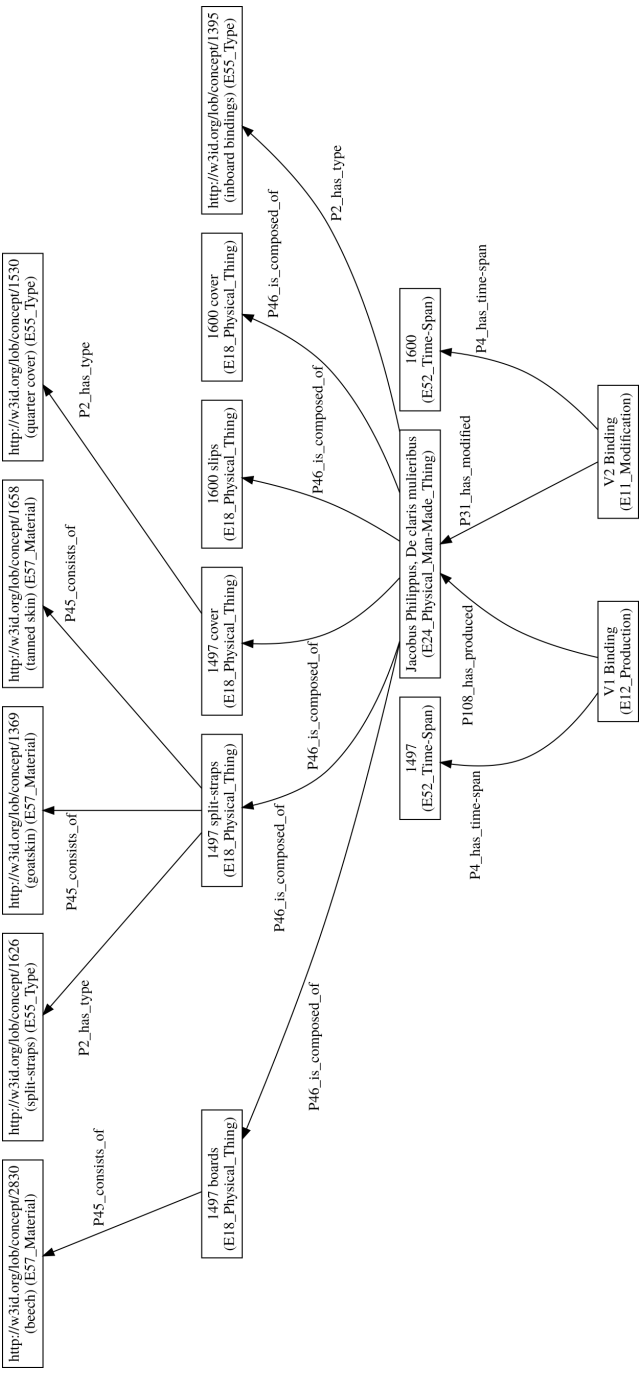


Figure 8: CIDOC CRM mapping of observations about the 2 versions of the binding.

book is *P110 augmented* and the property linking the event of adding the boards to the boards is *P111 added*.

We then consider an *E80 Part Removal* event, labelled as *V2 Board removal*, which happens during a longer modification event of the book around 1600. The properties of *P112 diminished* and *P113 removed* relate the removal event to the book and the boards respectively.

4.2 Cover

The description of the covers also involves the addition of the component during the first binding and its subsequent removal from the book. However, in this case we also have a second cover (*E18 Physical Thing*) added to the book as a replacement cover during the *V2 Cover addition* event. Both the *V2 Cover removal* and the *V2 Cover addition* occur during the longer modification event. To express the fact that one cover was removed before the next one was added we can use the property *P120 occurs before*.

4.3 Sewing supports

Another variation of this model is applicable to sewing supports. The split-strap sewing supports from 1497 were trimmed during the *V2 Binding* event. Trimming means cutting the slips at a specific length to match the thickness of the spine. The length of slip removed is not a separate entity prior to its trimming and therefore it may be difficult to argue that it is a *E18 Physical Thing*. Perhaps it is safer to consider the trimming of the slips as a more general *E11 Modification* event which occurs during the longer *V2 Binding* event.

An example of the output of this process encoded using the Resource Description Framework is presented in the Appendix.

5 Conclusion and discussion

In this paper we considered records as different versions of a binding using an event-centric approach. We encourage the production of records of events related to objects. There are two basic limitations of object-centric terminology when it comes to capturing the temporality of a component, namely: a) mistakenly grouping components from different periods/versions and b) lack of scalability. By switching to events we are able to describe any number of alterations/versions of components and we are able to separate components belonging to different versions.

Although we do not attempt to draw direct parallels with versioning tools in our discourse, adopting the principle of tracking changes is a useful model for describing

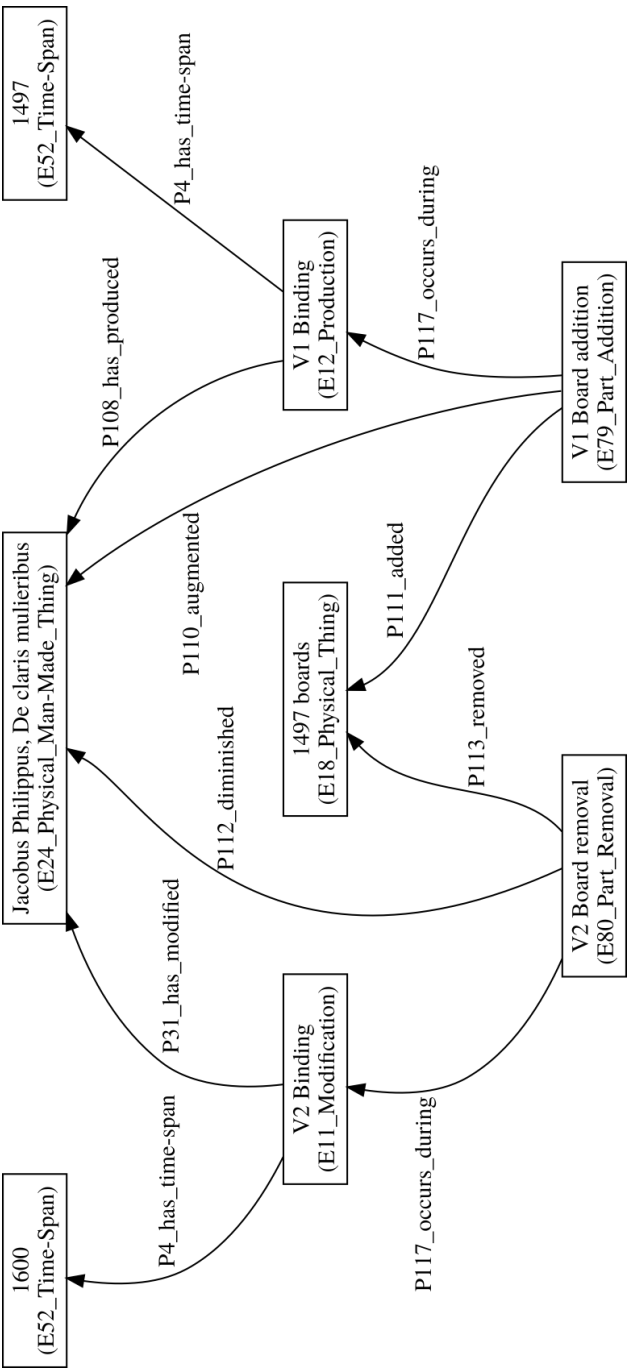


Figure 9: Detailed modelling of board components.

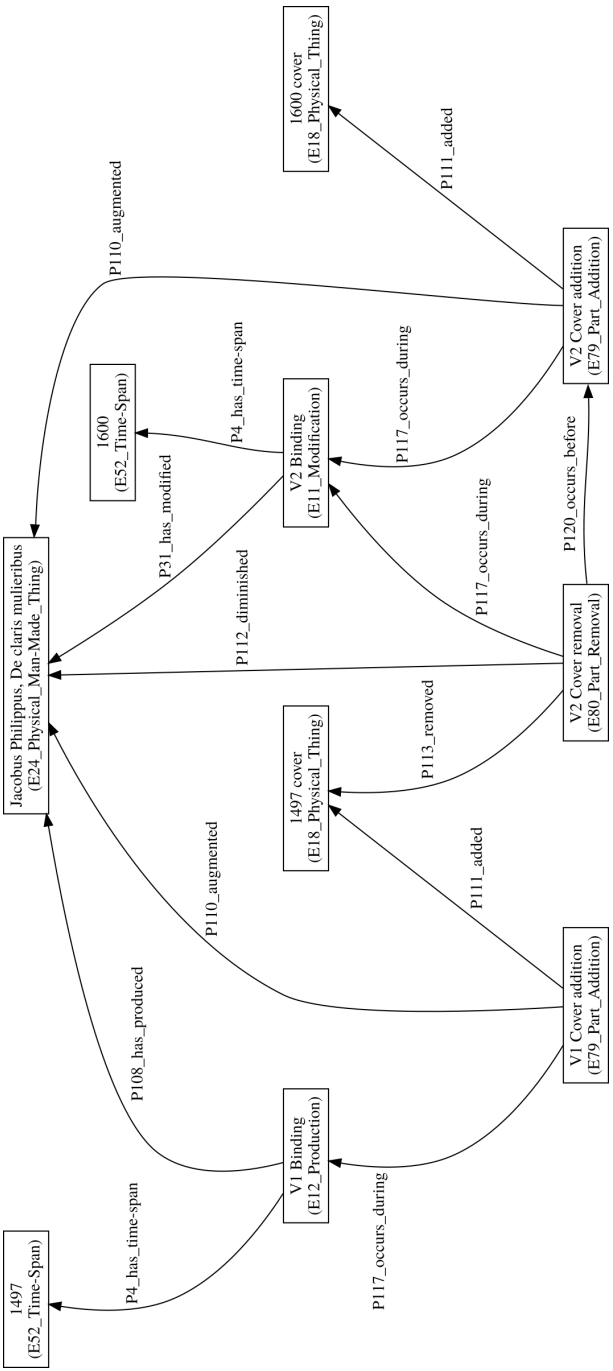


Figure 10: Detailed modelling of cover components.

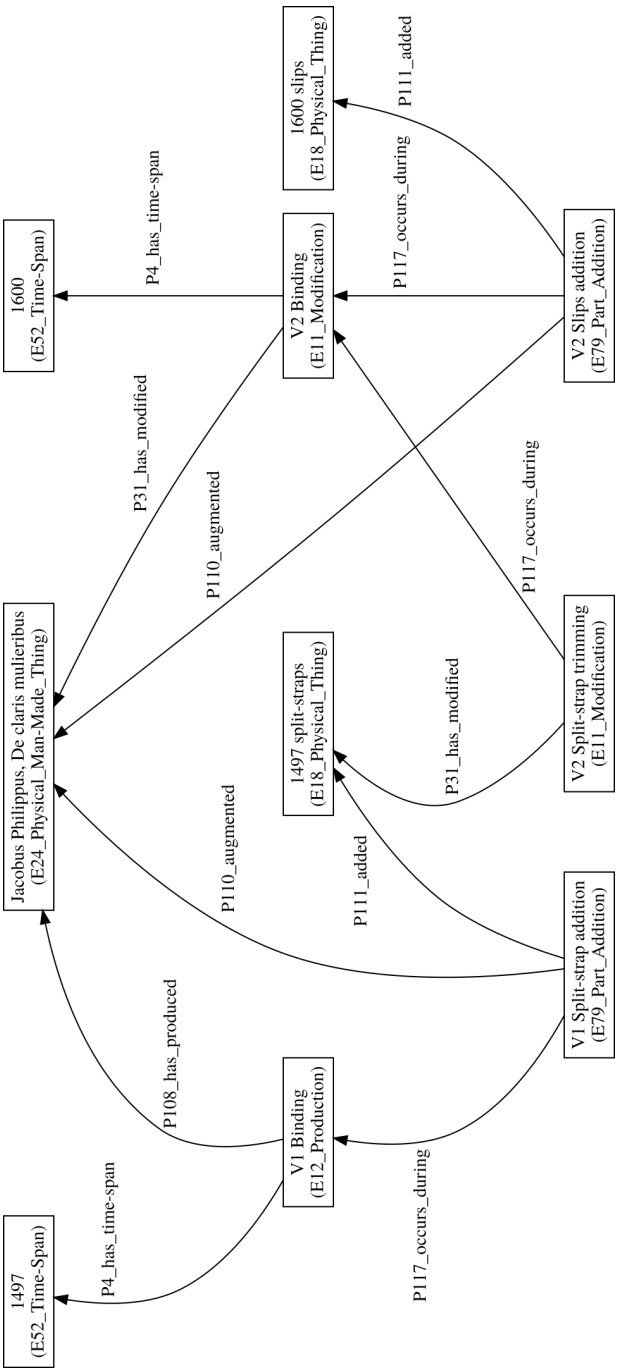


Figure 11: Detailed modelling of sewing supports.

the history of material objects. In the next sections we discuss some considerations which came up while modelling our case study.

5.1 Subjectivity

In this proposal we choose versions of the binding subjectively. Is it possible to be more objective about this choice? We think subjectivity is inherent in versioning. In computer programming it is up to the programmer to select the point when a new version of a file should be created. The choice of this point is subjective. In shared versioning systems there is an expectation that a committed change corresponds to a "bug-fix" or to the implementation of a new feature and therefore one could consider that these are more objective criteria for new versions. We can arguably apply the same principle to bindings. When re-attaching a torn leaf using *overcasting* (LoB, "Overcasting") or replacing a worn set of *endleaves* for the better protection of the *textblock*, a binder takes intentional action to fix the binding and perhaps this fix is a more objective criterion for setting new versions. Attaching a *bookmark* to an *endband* shows the need of marking the point in the text from which the reader needs to continue, therefore indicating a new feature of the binding. Perhaps new decorative or functional features are also valid objective criteria for setting new versions.

We do not intend to draw direct parallels between bookbinding history and programming but we are simply highlighting the wider issue of subjectivity in versioning.

5.2 Observation versus deduction

When experts survey bindings, they consider the evidence on the book under the prism of their experience. A sewing support which has been trimmed or broken at the joint may indicate the existence of longer slips and an earlier board or cover attachment. It is important to emphasise that the observation is only limited to the evidence on the book and that producing a record of the different versions of the binding is the result of deductive thinking based on training and previous experience. The proposed structure does not model any of these deductive processes. Because the records of the different versions of the object depend on these processes, perhaps a wider model to include inference methods should be considered. There is already extensive work in place to allow modelling and implementation of such a model (Doerr, et al.; Stead and Doerr).

5.3 Identifiers

The capacity of the CIDOC CRM model to scale according to the required detail of the resulting record means that in some cases a large number of identifiers need to be created to refer to each component and each modification event. In our case studies

we have used a simplistic set of identifiers but a large scale survey project including versioning records would need a clear strategy on the production of identifiers considering the following issues:

1. Persistence of identifiers: for how long would the identifiers need to be maintained and how would that affect migration to new systems?
2. Repeatability of production: how is it possible to reproduce the same identifiers for the considered entities in the future?
3. Human use: should human users (including developers) recognise entities by their identifiers?

5.4 Abstract schema

The abstract nature of the CIDOC CRM model may reflect our understanding of the world accurately but it may appear alien to the domain expert. For example, referring to part addition and part removal events is unusual language for the book conservator. Describing the replacement of the cover using a series of part addition and part removal events with multiple links to the book and the various components is not intuitive and there is significant amount of work to be done if documentation systems based on versioning and the CIDOC CRM are implemented for day to day work. It does, however, offer the possibility of recording complex data in a citable and structured way based on the observation of primary sources.

Bibliography

Arnim, Manfred. *Europäische Einbandkunst aus sechs Jahrhunderten: Beispiele aus der Bibliothek Otto Schäfer Schweinfurt.* Bibliothek Otto Schäfer, Schweinfurt, 1992.

Boal, G. "Conservation for Digitisation: Surveying the Arabic Manuscript Collection at the Wellcome Library." *The 5th Islamic Manuscript Conference.* Christ's College, University of Cambridge, 2009.

Campagnolo, Alberto. *Transforming structured descriptions to visual representations. An automated visualization of historical bookbinding structures.* 2015. University of the Arts London, PhD Dissertation. *UAL Research Online*, ualresearchonline.arts.ac.uk/8749/.

Darnton, Robert. "What is the History of Books?" *Daedalus*, vol. 111, no. 3, 1982. *JSTOR*, http://www.jstor.org/stable/20024803.

Doerr, Martin. "The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata." *AI Magazine*, vol. 24, no. 3, 2003, p. 75. www.aaai.org. Accessed 28 Mar. 2017.

Doerr, Martin, et al. "Factual Argumentation - a Core Model for Assertions Making." *J. Comput. Cult. Herit.*, vol. 3, no. 3, 2011, 8:1-8:34. *ACM Digital Library.*

ISO. *Information and Documentation: A Reference Ontology for the Interchange of Cultural Heritage Information.* ISO, 2006. International Standard ISO 21127.

Language of Bindings (LoB), edited by Nicholas Pickwoad and Athanasios Velios, Ligatus. www.ligatus.org.uk/lob/. Accessed 28 Mar. 2017.

McKenzie, D. F. *Bibliography and the Sociology of Texts*. Cambridge University Press, 1999. *ProQuest Ebook Central*, ezproxy-prd.bodleian.ox.ac.uk/login?url=http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=201933

Ravenberg, Heather. *A Data Model to Describe Book Conservation Treatment Activity*. MPhil. thesis, University of the Arts London, 2012.

RBMS Standards Committee. "RBMS Thesaurus." Association of College and Research Libraries (ACRL/ALA). www.rbms.info/vocabularies/binding/hierarchical_list.htm. Accessed 28 Mar. 2017.

Stead, Stephen, and Martin Doerr. *Version 0.7 | CRMinf*. Paveprime Ltd., 83.212.168.219/CRMinf/node/6. Accessed 29 Jan. 2016.

Stokes, Peter, et al. "Digital Resource and Database of Palaeography, Manuscripts and Diplomatic [DigiPal]." 2010, kclpure.kcl.ac.uk/portal/en/publications/digital-resource-and-database-of-palaeography-manuscripts-and-diplomatic-digipal(24c72efc-62d0-41f9-a2c9-bda3f4c09c81).html.

Velios, A., and N. Pickwoad. "Current Use and Future Development of the Database of the St. Catherine's Library Conservation Project". *The Paper Conservator*, vol. 29, 2005, pp. 39–53.

Appendix: Sample encoding in rdf/ttl

```
@prefix w3id:   <http://w3id.org/>.
@prefix rdf:    <http://www.w3.org/1999/02/22–rdf–syntax–ns#>.
@prefix xml:    <http://www.w3.org/XML/1998/namespace>.
@prefix rdfs:   <http://www.w3.org/2000/01/rdf–schema#>.
@prefix exa:    <http://example.org/>.
@prefix crm:    <http://www.cidoc–crm.org/cidoc–crm/>.
```

First binding event

```
exa:v1–binding a            crm:E12_Production;
   rdfs:label               "V1 Binding"@en;
   crm:P108_has_produced    exa:jacobus–philippus–de–claris–mulieribus;
   crm:P4_has_time–span     <uuid:AA>.

<uuid:AA>      a            crm:E52_Time–Span;
   crm:P82_at_some_time_within "1497"@en.

exa:v1–board–addition a     crm:E79_Part_Addition;
   rdfs:label               "V1 Board Addition"@en;
   crm:P110_augmented       exa:jacobus–philippus–de–claris–mulieribus;
   crm:P111_added           exa:1497–boards;
   crm:P117_occurs_during   exa:v1–binding.

exa:v1–cover–addition a     crm:E79_Part_Addition;
   rdfs:label               "V1 Cover Addition"@en;
   crm:P110_augmented       exa:jacobus–philippus–de–claris–mulieribus;
   crm:P111_added           exa:1497–cover;
   crm:P117_occurs_during   exa:v1–binding.
```

Second binding event

```
exa:v2–binding a            crm:E11_Modification;
   rdfs:label               "V2 Binding"@en;
   crm:P31_has_modified     exa:jacobus–philippus–de–claris–mulieribus;
   crm:P4_has_time–span     <uuid:AB>.


<uuid:AB> a                 crm:E52_Time–Span;
   crm:P82_at_some_time_within "1600"@en.

exa:v2–board–removal a      crm:E80_Part_Removal;
   rdfs:label               "V2 Board Removal"@en;
   crm:P112_diminished      exa:jacobus–philippus–de–claris–mulieribus;
   crm:P113_removed         exa:1497–boards;
   crm:P117_occurs_during   exa:v2–binding.

exa:v2–cover–removal a      crm:E80_Part_Removal;
   rdfs:label               "V2 Cover Removal"@en;
   crm:P112_diminished      exa:jacobus–philippus–de–claris–mulieribus;
   crm:P113_removed         exa:1497–cover;
   crm:P117_occurs_during   exa:v2–binding;
   crm:P120_occurs_before   exa:v2–cover–addition.

exa:v2–cover–addition a     crm:E79_Part_Addition;
   rdfs:label               "V2 Cover Addition"@en;
   crm:P110_augmented       exa:jacobus–philippus–de–claris–mulieribus;
   crm:P111_added           exa:1600–cover;
   crm:P117_occurs_during   exa:v2–binding;
   crm:P120i_occurs_after   exa:v2–cover–removal.
```

Book description

```
exa:jacobus–philippus–de–claris–mulieribus
                    a       crm:E24_Physical_Man–Made_Thing;
   rdfs:label               "Jacobus Philippus, De claris mulieribus"@en;
   crm:P2_has_type          <http://w3id.org/lob/concept/1395>;
   crm:P46_is_composed_of     exa:1600–slips, exa:1497–covers,
                              exa:1497–split–straps, exa:1497–boards.
exa:1497–cover      a       crm:E18_Physical_Thing;
   rdfs:label               "1497 cover"@en.

exa:1497–split–straps
                    a      crm:E18_Physical_Thing;
   rdfs:label               "1497 split–straps"@en;
   crm:P2_has_type          <http://w3id.org/lob/concept/1626>;
   crm:P45_consists_of      <http://w3id.org/lob/concept/1658>,
                            <http://w3id.org/lob/concept/1369>.

exa:1497–boards     a      crm:E18_Physical_Thing;
   rdfs:label               "1497 boards"@en;
   crm:P2_has_type          <http://w3id.org/lob/concept/1222>;
   crm:P45_consists_of      <http://w3id.org/lob/concept/2830>.

exa:1497–covers     a      crm:E18_Physical_Thing;
   rdfs:label               "1497 covers"@en;
   crm:P2_has_type          <http://w3id.org/lob/concept/1530>.
```

```
exa:1600_cover     a     crm:E18_Physical_Thing;
   rdfs:label            "1600 cover"@en.

exa:1600_slips     a     crm:E18_Physical_Thing;
   rdfs:label            "1600 slips"@en.
```

Types from thesaurus terms

```
<http://w3id.org/lob/concept/1658>
        a               crm:E57_Material;
        rdfs:label      "tanned_skin"@en.

<http://w3id.org/lob/concept/1369>
        a               crm:E57_Material;
        rdfs:label      "goatskin"@en.

<http://w3id.org/lob/concept/2830>
        a               crm:E57_Material;
        rdfs:label      "beech"@en.

<http://w3id.org/lob/concept/1395>
        a               crm:E55_Type;
        rdfs:label      "inboard bindings"@en.

<http://w3id.org/lob/concept/1626>
        a               crm:E55_Type;
        rdfs:label      "split_straps"@en.

<http://w3id.org/lob/concept/1530>
        a               crm:E55_Type;
        rdfs:label      "quarter covers"@en.

<http://w3id.org/lob/concept/1222>
        a               crm:E55_Type;
        rdfs:label      "boards"@en.
```

Versioning Charters: On the Multiple Identities of Historical Legal Documents and their Digital Representation

Georg Vogeler

Abstract

This chapter proposes a model for the concept of *versions* and how it can be applied in the scholarly discipline of diplomatics, the study of historical legal documents. It describes the various concepts and physical things the discipline of diplomatics connects with the term *charter*, as well as the practice of people working with them. The chapter also connects the history of preparing, engrossing and copying charters, with the archival and scholarly practices of describing, editing, or photographing, including transforming charters into digital representations.

By drawing on the Functional Requirements for Bibliographical Records (FRBR), the Vocabulaire Internationale de la Diplomatique, and charter databases such as monasterium.net and *The Making of Charlemagne's Europe*, the author argues that a model for *versions of charters* should not start with a definition of *charter*, but rather with the network of relationships which can be considered instantiations of *versioning*. W3C Resource Description Framework (RDF) representations of the data fragments used to represent a charter—for example images, descriptions, texts, legal actions, archival and other identifiers—allow a giant graph of charter versions to be created and help to use and approach the rich set of charter databases as integrated resource.

1 Introduction

This chapter proposes a model for the concept of *versions* and how it can be applied to the digital representation of charters, historical legal documents. These representations are often stored in, and published as, databases, which arguably seems to be the most appropriate method for this kind of cultural heritage (Vogeler, "Digitale Urkundenbücher"). The largest of these databases is likely the monasterium.net portal. It contains more than 600,000 charters, and will be used as an source of examples in the following chapter. Besides monasterium.net, there are many other rich databases for charters. The following list names some of the most prominent examples from the rich variety available online:

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 127–150.

- Chartae Burgundiae Medii Aevi (Projet CBMA)[1] (Magnani - Gasse-Grandjean "CBMA Les débuts du projet," "CBMA, part I-V "; Rosé),
- Diplomata Belgica (Hemptinne et al.; Deploige et al.)
- ArQuibanc[2] (Piñol Alabart)
- DEEDS[3] (Gervers, *DEEDS*; Gervers, et al. "The Deeds Database"; Gervers and Margolin, "The Deeds Project," "Managing Meta-data")
- The making of Charlemange's Europe database (Rio et al.)
- Cartago (Stichting Digitaal Oorkondenboek Groningen en Drenthe; Heidecker)
- A database of original charters for Germany issued before 1250 (Institut für Mittelalterliche Geschichte der Universität Marburg; Bischoff "Die Datenbank"; Roberg and Klipsch; Baumbach and Meyer)
- Pergamene di Puglia online[4]

All of these resources try to assemble information on charters from different sources, and often, the same charter is published several times in different places simultaneously. For example: the documents recorded in Felix Henri d'Hoop (1870) are held on both monasterium.net[5] and the Diplomata Belgica.[6] Similarly, many of the charters recorded in the Charlemagne-database can also be found in monasterium.net.

For instance, the entry charlemagneseurope.ac.uk/browse/charters/415/ refers to the same charter as monasterium.net/mom/DE-HStAMa/UrkHersfeld/2254/charter, a diploma issued by Charlemagne to the abbey of Hersfeld in 775 (MGH D Kar 89). While the Charlemagne database gives a highly structured description of the transaction recorded in the charter, monasterium.net provides digitised images from the archives. As the monasterium.net portal aggregates metadata from archives and printed editions, much of the metadata might be duplicated in different places on the internet. For example, libraries might have put printed descriptions of the charters online, whilst many of the charters can be found on the website of the archive as well. In order to reconcile all of these sources of information, it is necessary to construct a thorough data model for the relationships between all of the various digital charter representations. This paper considers these relationships as specialisations of the relationship *versions of*, which is the major concern of this volume.

Over recent years in digital humanities, the W3C proposal for a *semantic web* (Berners-Lee, Hendler and Lassila; W3C *Semantic Web Activities*) has become the

---

1 www.cbma-project.eu/.
2 www.ub.edu/arquibanc/.
3 deeds.library.utoronto.ca/.
4 www.sapuglia.it/index.php?option=com_content&view=article&id=213&Itemid=214.
5 monasterium.net/mom/SaintBertin/collection.
6 e.g. www.diplomata-belgica.be/charter_details_en.php?dibe_id=2952, is the same charter as monasterium.net/mom/SaintBertin/3e0efb41-2ece-4e01-bb9f-69f2437ec7a7/charter.

go-to reference method for publishing structured data on the web.[7] This technology allows a single charter to be uniquely identified on the internet by assigning a uniform resource identifier (URI) to it. Other databases can then refer to this URI, for example by properties such as `sameAs`, defined in the W3C semantic web standard *OWL* (W3C *Web Ontology Language*), or as `exactMatch`, defined in the W3C semantic web standard *SKOS* (W3C *Simple Knowledge Organisation System*; Miles and Bechhofer; Isaac and Summers). However, this is only possible if the charter databases agree on the ontological question, "what is a charter?"

The following will attempt to demonstrate that the study of diplomatics has no clear answer to this question, instead offering a rich set of various different concepts for *charters*, represented in charter databases. Additionally, there are further concepts in data models for these databases, for items which could be considered *versions* of charters. Because of this, a semantic web data model for charters is necessary to consider both in tandem: the complex and various meanings of the term charter, and the fact that different versions—both physical and digital representation—of charters exist.

These considerations start with an outline of the concept of *charters* as legal documents, in particular those from the European Middle Ages and early modern times. Firstly, it will be more clearly established what diplomatics—as a well-established historical auxiliary science—considers charters and which concepts have been developed through scholarship to describe different versions of the charter. Following this, it will be discussed to what extent the term version is useful, or if other terms like *description*, *representation*, *surrogate*, *revision*, *adaption*, or *instantiation* can help to provide a clearer picture. The various relationships added by digital technologies will be presented, followed by a final proposal for a conceptual model for *versioning charters* which could be expressed in RDFS.

2 Scholarship

2.1 Basic concepts

Charters are good examples with which to highlight the complexity around the versioning of cultural objects. This is because of their physical and textual form, and their relation to underlying concepts and social activities involved in their creation and use. This is already the case when studying the historical practice connected to charters, and when studying the process of their digitisation. The complexity of the problem becomes clear when leaving the assumption that charters are just a specific form of text. They are, in fact, much more: in this chapter the term "charter"

[7] In 2013 the W3C moved the Semantic Web activities into a newly founded *W3C Data Activity*.

is used to relate to a core concept for work undertaken in the scholarly discipline of diplomatics.[8] It is not used as in Medieval Latin (du Cange, II, 292a s.v. *Charta* (1)), but in the widest possible sense, wider even than the definition provided in the Vocabulaire Internationale de la Diplomatique (Carcel Ortí, in the following: VID). Because of this, it has become similar in use to the term *document*, which has gained a much wider meaning due to its use in modern office technologies, where every bit-stream representing something readable by humans can be called document.

There are several English terms which have a similar scope and meaning, or can at least be considered specialised forms of charters, such as *deed*, *instrument*, *title*, *written document*, *act*, *record*, and *indenture*. This broad interpretation is close to the French tradition, which considers all archival documentation to be subjects of the field of diplomatics; an approach followed by Leonard Boyle in his definition of this area of study. Still, this interpretation has not become common in the community of diplomatists (Kölzer).

Even more recent English publications on charters focus on the judicial value of the charters (Mostert and Barnwell; Jarrett and McKinley), although they extend the scope of diplomatics into cultural history. Therefore, applying the term *charter* in the context of diplomatics gives four concepts which form part of the core definition: A charter is something written (1), which gives evidence (2) of a legal fact (3) by means of formal properties (4), which are stable for a specific time period and geographical area. This is a rough translation of the classical definition of *Urkunde*, given by Harry Bresslau at the turn of the 19th century, as a form of *summa* from the golden age of diplomatics:

> *Urkunden sind schriftliche, unter Beobachtung bestimmter, wenn auch nach der Verschiedenheit von Person, Ort, Zeit und Sache wechselnder Formen aufgezeichnete Erklärungen, die bestimmt sind, als Zeugnis über Vorgänge rechtlicher Natur zu dienen.* (Harry Bresslau, Handbuch der Urkundenlehre, 2nd ed., 1915, p. 1)
>
> Charters are written declarations recorded in compliance with certain forms, alternating according to differences in person, place, time, and matter, which are meant to serve as a testimony of proceedings of a legal nature (my translation)

The definition in the already-mentioned VID (Carcel Ortí) follows these lines:

> *Les* ***sources diplomatique*** *forme d'une part, des actes écrits; de l'autre, des documents résultant des actions juridiques et des activités administratives et*

[8] The most recent resumé of the scope and history of diplomatics is given by Theo Kölzer.

> *financières de toute personne physique ou morale; enfin des lettres expédiées ex-officio et dont la forme est soumise à certaines règles.*
>
> The diplomatic sources are first: written acts; second: those documents which result from legal acts, and administrative and financial activities carried out by any kind of physical person or legal body; and third: official letters which therefore have a form following certain rules. (my translation).

This is in line with the use of the word *charter* in the English language since the Middle Ages, as it is documented in Glasgow's *Historical Thesaurus of English* (Samuels et al., s.v. "charter"). Therefore, a working definition for this paper, which deals with different perspectives on the subject, might be that charters are written documentations of legal acts in their historical development. This includes testaments, wills, contracts, privileges, orders, obligations, certifications, and similar. In this chapter, European medieval charters are used as the primary example of charters. Certainly, the tradition has roots in Roman administration and legal culture, and it also has followers in early modern times. I have argued previously that the concepts developed in European medieval documentation, the connected conceptual models, and the technical realisations in formal ontologies and schemata could even be applied globally (Vogeler "Digital Diplomatics").

The working definition and the definition given by the VID already lead to the first important concept, which must be considered when talking about versions of charters. Diplomatists study the charter as a double instantiation:

- the legal act executed by humans or the legal fact accepted by humans in the past
- the artefact created by humans to document this act or to bring this fact into existence

This difference is discussed in much of the recent scholarship around diplomatics (e.g. Heidecker; Mosert and Barnwell; Jarrett and McKinley; Barret, Stutzmann and Vogeler), which studies charters as records which "owe their existence to the fact, that there were people at one time or another who had felt the need to […] preserve in writing the memory of a transaction or event" (Boyle, 89). Research could therefore profit from a clear modelling of this double instantiation.

This distinction between artefact and abstract legal fact is similar to the relationship between the FRBR concepts of *work* and *item* (IFLA; Bekiari et al.). Consequently, written artefacts could be considered embodiments of one abstract work, whilst at the same time each being considered a version of the other.

It becomes complex when we want to talk more specifically about the relationship between the written artefacts. These versions can be classified according to how they instantiate the legal fact. Diplomatics has developed a detailed set of considerations for this relationship. There is a legal perspective, in which the relationship can be

distinguished between artefacts which create the legal fact (dispositive document, *charta*)—whose destruction eliminates the legal fact (Sennis)—and those which document an existing legal fact in a way that it can be used as proof in court (evidential / probatory document). In a temporal perspective, the instantiations of the legal fact can represent different stages in the production and use of a charter:

- The *engrossment* or *original* (VID 42) is the version which legal discussion would refer to as authoritative
- A draft can precede it, which is a non-accorded preparatory text
- Finally, copies can follow an engrossment, which convey the correct text representing the legal fact, but carry legal value only in their reference to the original

Going deeper into the historical documentary practice, further forms may be distinguished. In early medieval times in many regions north of the Alps, charters were only considered a written means for memorising a transaction, and the people who could testify this (Johanek; Molitor, "Das Traditionsbuch," "Zum Traditionsbuchwesen"; Härtel 108–17). They are thus evidential, but lack any intrinsic legal value themselves. These *notitiae* were written in a less formal way, sometimes in preparation for, or during, the ritual which brought the legal fact into existence. Many of them have only survived in books (*libri traditionum*), where they were stored in order to gain an overview of monastic possessions and to create a collective memory of the relationship between benefactors and monks (Borgolte). In this case, there is no engrossment of the legal act.

Furthermore, diplomatic culture has created other forms of valid written documentation that should be considered in the context of versioning: in northwestern Europe and England in the 9th to 13th centuries, the practice of *indenture* (or *chirograph*) was widespread. The parties of a contract wrote a duplicate of the text on one parchment, cut it in half between the two texts—often through a word like *chirographum* or through the alphabet written in this place—with each party handed one part to preserve (Bischoff, "Zur Frühgeschichte"; Trusen; Parisse; López and Encarnación; Lowe; Herold; Bedos-Rezak; Groß). Each part could gain value as proof when it corresponded and matched the other. This created two written artefacts, which—theoretically—only represent the legal fact when viewed together, although in historical practice each single part served as documentation in court.

An additional fact to consider is that the diplomatic practice over time creates versions of the same legal fact in different wordings. That is obviously the case with translations, but it even happened in a culture in which no neutral form of contractual agreement existed, so each party had to create a charter declaring its own will to agree on the legal fact. The agreements between the city and Bishop of Lübeck between 1220 and 1230 demonstrate the variety of diplomatic forms that this could take: for example, charters issued by third parties, two charters with the same text and the

same seal but with different names as issuers, and charter text in two engrossments with two seals (Prange).

The Papal administration is famous for similar cases of double documentation. When rights were granted to petitioners, one document was issued addressing the beneficiary, and a second was issued ordering a close by ecclesiastical official to execute the grant. The two documents are of course similar in the core legal descriptive text, but differ in the manner of communication with the addressee. They also differ in form, as the grant carries more solemn features than the order. The difference in physical features has even led to a terminology for these documents. The grant is called *littera cum serico*, as the thread connecting the seal to the parchment is made of silk, while the order is called *littera cum filo canapis*, after the hemp used for its thread. This practice of double documentation even creates single entries for both charters in modern archival and diplomatics metadata (e.g. Barbiche no. 296–no. 300).

This combination of both grant and order were practiced in other administrations, such as that of the Normans in the Kingdom of Sicily (Enzensberger 98–100). Considering the two documents as the representation of a single legal fact hides that the two were designed for entirely different social interactions (order and permission), so the individual legal fact could be considered as a version of the common legal fact. Administrative practice in the Middle Ages, as well as archival practice, shows that this was a common approach. It was usual for the beneficiary to receive both pieces of parchment: the one carrying the text with the grant and the one ordering the grant to be executed. From the grantee's perspective, both documented the same legal fact.

The clerks also created notes when preparing formal engrossments, for example on the back or in the margin of the document. In the Middle Ages, the papal chancery was the first to establish note taking as part of the procedures in central administration, and other central administrations followed (Csendes et al.). In Italy, a different notarial culture developed in the 11th century and spread over the whole continent in the centuries that followed. In this culture, the notary was a person involved in the documentation of a transaction to secure a neutral and authentic version of the agreement. His credibility was so strong, that the notes of the transaction in a notary's register (the *imbreviatura*) could be accepted as proof in court (Costamagna 22–4; Härtel 83–7).

Administrative and legal culture created a variety of other forms of copies, for example those collected in chartularies, those copied on single sheets with no further context, and those inserted into historiographic narratives. The colloquium of the Commission Internationale de la Diplomatique in 1999 (Kosto and Winroth) studied examples of these and even the copies of charters can take a variety of forms, meaning that the legal fact can be taken over by the issuer of a new charter.

This is particularly the case when sovereigns inherit the throne, and older rights are confirmed as still existing. They refer to the *retroacta*, i.e. the charters of the

predecessors brought to the prince's court to prove the rights of the petitioner. At least in German diplomatics, this type of copy is also known as *transsumptum*, while copies executed by people claiming not to be involved in the legal fact carry the name *vidimus*. British diplomatics uses for the latter example the term *inspeximus*. In the case of the *vidimus* the original presented to the court or to the notary is only copied verbally, claiming that the copy and the original are verbally identical and thus prove the same fact. The last method was often used to create formally-incontestable versions of forged documents. Both types of copies repeat large parts of the text of the original, thus they are subsumed under the term of *insertum* in diplomatics terminology, which again creates an unclear situation when talking about versions of charters and leaves a number of questions unanswered. Is it right talk about the full charter, including framing text and inserted original? Or do we talk about the inserted text taken from the older document? Is the copy just a version of the original or an original in its own right, citing the text of an older charter?

The variety of different versions of charters can probably best be modelled starting with the relationship between legal fact and written artefact. The legal fact can be considered a common reference point. Versions of this take the form of written artefacts which are used for a number of different purposes. Some of them can be used as proof in court, some of them bring the legal fact into existence, and others are just a support for memorising information. The legal fact stated in the charters might never have existed (forgery) or the wording given might have changed according to the textual form of the written documentation.

The FRBR term *manifestation* for this kind of relationship might be helpful to reduce the term *version* to relationships between these manifestations of one legal fact/act only. Diplomatics terminology offers different typologies for these versions as *draft*, *imbreviatura*, *engrossment*/*original*, *authentic copy*, *copy*, *multiple exemplars*, *duplicate*, *vidimus*, or *transsumptum*. For most of them, an accepted VID definition exists, and the concepts can be represented in SKOS in the following way (Vogeler, "Von der Terminologie"):

- vid:353 for the draft[9]
- vid:357 for the imbreviatura
- vid:46 for the engrossment
- vid:54 for an authentic copy
- vid:53 for any kind of copy
- vid:43 for multiple exemplars, to which
- vid:45 (duplicate) is a specialisation.

Only the distinction between *vidimus* and *transsumptum* is defined differently by the VID, as noted by Rolf Große in 1996.

[9] The prefix "vid" stands in for the namespace string, "www.cei.lmu.de/VID/#VID_".

All of these manifestations of the legal act demonstrate that the sequential relationship which they may possess has to be considered independent from the legal status. While the duplicate originals in a chirograph are contemporary to each other, copies are created later and imply the existence of an original as antegraph. A draft, or the *imbreviatura*, precedes the engrossment and suggests its existence (although it might have never existed). The legal culture around the *notitiae* allows for multiple non-contemporary versions, which can simply be other manifestations of the legal act, while no legally binding original was ever produced. The basic concepts behind the sequence are therefore not very well covered by the terminology of diplomatics itself and could be reduced to the relationships between antegraph, apograph, or duplicate. This certainly applies as well to copies of copies, which leads into the administrative, archival and scholarly practice of creating new versions of a charter in later periods.

2.2 Handling the tradition

The versions of a charter created at a substantially later point than the legal act are handled in several different distinct communities, including administrative practice, archives, and scholarship. Their individual approaches to the charters create other types of versions. It seems straightforward to cover these by using terms such as *description*, *metadata*, *representation*, or *surrogate*, suggesting that they are only referencing the original. Facsimiles are considered surrogates, archival metadata would be called description and scholarly edition would be classified under representation. These forms are compatible with the historical practice described above: copies are the results of the administrative practice in the same way as archives. Thus, archivists worked like medieval copyists and created subject-oriented collections, sorted by subject or issuer. This change in context adds information to the single document and can therefore be considered a new version. Since the 19th century, archivists have changed their approach and now consider the artefacts part of historical records. Most of them follow the archival principal of *respect des fonds* / *Provenienzprinzip*, established as best practice during the 19th century (Mueller et al. 1898; Schwineköper; Uhl). This meant that charters had to be put back into the context from which they originated, again changing the context and therefore the interpretation of the charter. There is even a discussion around whether charters would require a different way for the principle of respect des fonds to be applied (Hartmann and Engelhardt).

Nevertheless, most of archival practice is well covered by the term *description*, which involves metadata helping the archivist to handle the artefacts and the historical researcher to find information documented by the charters. Putting this combination of features of the written artefact together with a verbal description of the legal fact in relationship to other versions can create confusion: There are archives which

prefer the content perspective, putting two physical objects in one description. The Papal chancery issued the incorporation of Berchtesgaden into the archbishopric of Salzburg on June 16th 1393 (AUR 1393 VI 16) in two verbally-identical charters, both authenticated by a Papal bull. The archivists in the Haus-, Hof- und Staatsarchiv decided to put both pieces into one metadata entry.[10]

Focusing on the legal act as well, other archivists split the description of one single artefact into two entries, to reflect the multiple legal facts reported. This can be found, for instance, in the copy of a document for the Hungarian King Andreas II by the chapter of Bratislava in the National Archives of Slovakia, which is in two entries: one for the copied charter (n. 64 ins. 1.1) and one for the charter copy (n. 64).[11]

Scholarship has developed its own methods of representing charters, and they bring another way of conceptualising charters to light. In print culture at least, scholarly editions are considered good representations of a charter. Typical scholarly editions of charters demonstrate that a charter is a combined object. For example, modern editions like those in the MGH Diplomata series include a verbal description of the legal content (*regest*); the transcription or critical text of the document; a description of the textual witnesses to the document; and a critical comment reflecting on the authenticity status, the production and the historical context of the document. It therefore represents all facets of the charters which have been discussed in the first section of this paper: the legal fact (in the *regest* and the critical comment), the artefacts carrying a text (in description of the textual witnesses and the very text itself) and the relationship between both in the critical comment.

However, this also provides a further representation, namely the abstract "text" as reconstructed in a stemmatologic critical edition. Michele Ansani (2006) argued that this method is better-adapted to the study of charters than it might be to other medieval texts. With charters it can be assumed that one authoritative original existed from which all copies derived in different ways. It can also be assumed that the existence of the original was implied in copies—at least in authenticated ones—and most likely in forgeries which gain impact only by being assumed as original. Literary texts on the other hand might result from oral traditions, which were simultaneously written down in different versions, and indeed gained only presence in contemporary culture if the single manuscript was read, which was the authoritative version to the reader or listener.

In the beginning of this chapter the types of charters were introduced, to which Ansani's assumption does not apply (*notitiae*, duplicates etc.), but his position still holds true in the work of 19th and early 20th c. scholarship where the text of a charter was a separately-existing item. FRBR can help to understand this better when it sets

[10] monasterium.net/mom/AT-HHStA/SbgE/AUR_1393_VI_16/charter.

[11] monasterium.net/mom/SK-SNA/4156-SukromnyArchivBratislavskejKapituly/64%28ins_1.1%29/charter and monasterium.net/mom/SK-SNA/4156-SukromnyArchivBratislavskejKapituly/64/charter.

the expression level as essential for bringing the abstract work into existence, while it still does not have to be physically embodied in a manifestation or an item. The stemmatological scholarly editions consider this abstract concept as *text*.

The focus on the text of a charter as an abstract object leads to another form of a charter itself. The linguistic skills necessary to understand the original text of the charters cannot be expected from modern students. In the European Middle Ages most of them were written in Latin, and even vernacular texts are often not any easier to understand for modern students. As charters are an important source of historical information, it seems that translations into modern languages are needed to provide access to the content of the charters—this creates another type of version to consider. An example in print is the source collection in the Freiherr-vom-Stein-Gedächtnisausgabe (Buchner and Schmale). Digital examples of this are the results of a teaching experiment undertaken by Tilmann Lohse in Berlin. Even contemporaries created translated duplicates of a charter (Schulze).

Like historical copies and archival descriptions, scholarly editions can create several different representations of a charter just by re-contextualisation: charters published in a regional collection (Kölzer et al.) can get into a scholarly edition organised by issuer or by archival fonds. This does not change the physical description or the textual representation but can alter the description of the content. Abstracts can highlight information of more importance in one particular context. They can even reduce the content of a charter to partial information of relevance in a totally new context.

The printed version of the Chartularium Sangallense (Clavadetscher and Sonderegger) is an example of this—and with it the online version on monasterium.net. The Chartularium Sangallense contains full editions of all charters if the author, addressee, or the subject is from the Canton of Saint Gall. Additionally, it records all other charters mentioning persons from the region as abstracts highlighting this person. For example, the charter by the Provost of the Cathedral in Zurich confirming the endowment of an annual Mass in the year 1327 is linked to the Canton of Saint Gall only by the provost's origins in Toggenburg, in the heart of the Canton (Clavadestscher and Sonderegger vol. 6, n. 3307). The legal fact reported by these charters might be similar to all the others, but for the research interest of the editors of the Chartularium Sangallense, the name of one witness is more important than the possession granted by the Emperor to a third party.

In addition to the versioning of a charter as draft, engrossment and copy, or as expressions and manifestations of the legal fact, scholarship and archival practice creates additional versions of charters. Examples of this include calendars like the Regesta Imperii or Saywers list of Anglo-Saxon Charters; scholarly editions such as the Monumenta Germaniae Historica or the British Academy Anglo-Saxon Charters series (Campbell et al.); and archival descriptions.

The abstract in a scholarly calendar represents the same charter as the full edition, in the same way that metadata created by the archives does. However, all of them reflect different properties and interests in the charter. For example, the calendar and archival abstracts refer to the legal fact or to historical facts; critical editions represent an abstract text which is a reconstruction based on the relationship of the textual witnesses or analysis of external features and archival conservation work with the artefacts.

The question therefore arises, around whether the relationship between abstract *work* and any forms of expression and embodiment, as suggested by the FRBR model, should really be applied to the relationship between legal fact and written artefact. On the contrary, it seems appropriate to conceptualise the *charter* as an abstract concept on the FRBR work level. This concept refers to an activity of people in the past through which they tried to establish a specific personal relationship with strong bindings, or *legal fact.* The abstract concept of a charter would then be defined by the possibility to find an expression and a physical embodiment of this legal fact. Indeed, many charters are only known by reference in other documents or historiographical reports, a concept which the German diplomatic scholarship calls *deperditum.* Consequently, this would suggest that the concept of *charters* should be defined as a *possibility* rather than an actual *work* according to FRBR. Following the FRBR model, the major form of *expression* is the text of the charter, although it should be taken into account that documents usually carry physical or visual features, such as graphical signs, signatures or seals, which express an important part of the legal fact, of which the linguistic text is not a sufficient *expression.*

3 The digital world

Transferring all these different perspectives on the concept of *charters* into the digital world creates another layer of versions: Certainly, there are the digital transformations of older forms, usually as XML data as they are considered structured text and the use of digital photography has added a visual surrogate to the descriptions and transcriptions.

Beyond the core study of diplomatics, another form of digital representation emerges, which is based on the legal fact documented by the charter, namely that the content of the documents is transformed into databases relating to various research interests. For example, prosopographical databases allow the study of personal networks and careers. Geographical information from the place of issue, the recipient or the location of property allows itineraries to be reconstructed, leading to a definition of a region as *königsnah* or *königsfern*—the concept developed by Theodore Mayer has since been frequently used to interpret regional power constellations—and insights to

be gathered around the distribution of demesne. Since such a database uses a set of information from the charter, every charter entry in the databases can be considered a separate version and calendars are reduced to the facts of interest in the database. The charter itself gains the ontological status of a *source* of information.

The ability of digital media to be easily modified makes this even more complicated. Gunter Vasold describes how the scholarly practice of the division of labour, of revision, and of re-contextualisation could be converted into the digital world, all involving modification. The working group around Ray Siemens calls a part of this practice *social edition*, namely: that a community of practice uses the modern online tools for collaboration on a scholarly edition. This can be done by involving volunteers to help with transcription, by publishing user comments, or by using collaborative bibliographic and text creation tools, for example (Siemens et al.). All of them demonstrate that the digital representation of a charter is not stable. Any model of versions of charters therefore has to take into account the multiple possibilities created by digital versions. As these versions are part of scholarly practice, they can be considered as interpretations or as translations into current discourse, allowing them to remain meaningful or for further meaning to be attributed to them.

4 Formalisation of the model

Figure 1 attempts to visualise the theoretical result of the considerations above. The *charter* frame in figure 1 describes the area in which the relationships between all the concepts considered a *charter* converge. Many of them point to each other, but it is not clear which one is *the charter*. Lots of them could be considered to be instantiations of the written artefact (*draft*, *engrossment: original*, *notita*, *copy*), and a set of these could be used in court (*imbreviatura*, *engrossement: charta*, *authentic copies*). The *charter* frame is easier to identify by the conceptualisations pointing from the inside of the frame to the outside, such as the historical fact documented by the *charter*, or by those pointing form the outside into the frame, such as the digital representations of the *charter*. However, in practice, many of the outside concepts refer to only single concepts in the core of the *charter* area.

4.1 Serialising the model

Developing a consistent model for the versioning of charters has high relevance in the development of a charter portal such as monasterium.net, the world's largest portal for medieval and early modern charters. The source for this material is usually archival data, but it also contains 5,348 transcriptions from the DEEDS dataset,[12]

[12] monasterium.net/mom/DEEDS/collection.

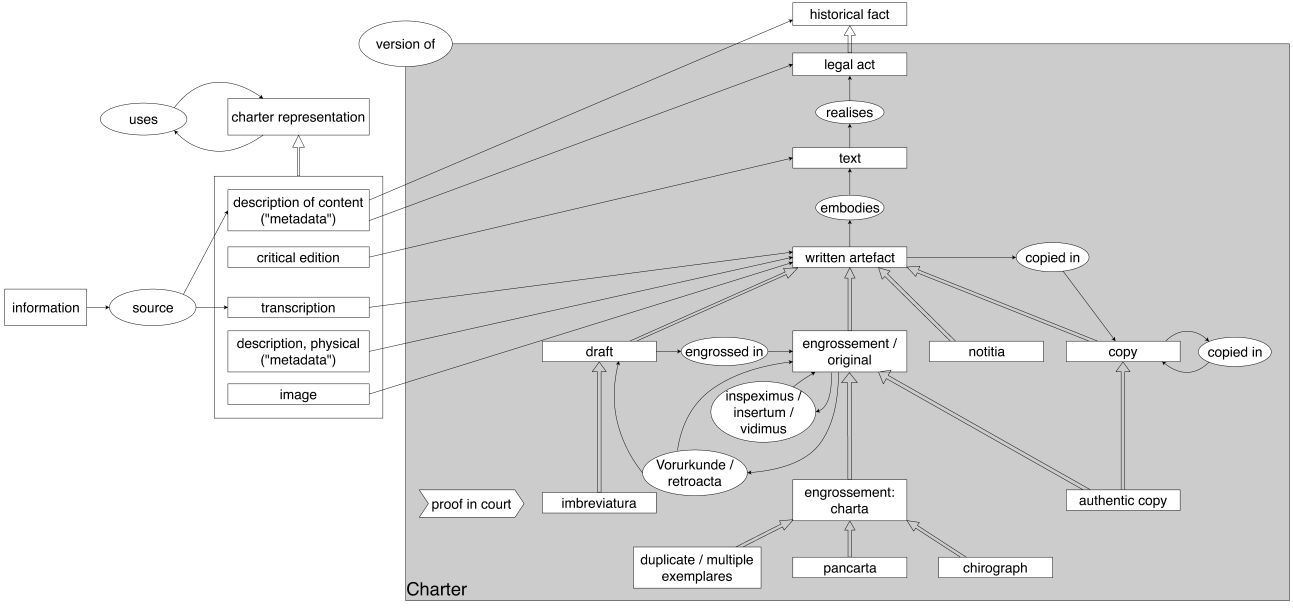


Figure 1: Graphic of concepts which could be considered versions of charters (and some related concepts).

and more than 40,000 charters extracted automatically from Google OCR.[13] Several archives only provide links to the image hosted on their own servers (for example, the Florence State Archives[14]). Others provide images, but almost no metadata, like many charters from the archives of the bishopric of Passau.[15] Currently, monasterium.net is ingesting data from the Regesta Imperii Calendar (Rübsamen; Regesta imperii V,1,1[16]). Furthermore, projects use it as a platform to publish research collections, such as the *Illuminated charters* project (Roland et al.), which studies charters bearing images or rich decoration from all over Europe.

This variety shows that it should be vital to have a consistent concept of versioning in the resource. The least problematic case is the versioning of the single entries—each charter is represented by an XML file and every change of this file made public can be stored with the versioning functionality built into the native XML-database in the backend (eXist-db). Some of the relationships developed above are part of the data model of each single charter.

The XML schema used is based on the Charters Encoding Initiative (CEI)[17] and is available on GitHub.[18] This schema has sections for the description of the content (`cei:chDesc` with `cei:abstract`, `cei:issued`, `cei:issuer`, and `cei:recipient`) and additional keywords can be marked up as `cei:persName`, `cei:placeName`, `cei:geogName`, `cei:organisation`, or `cei:index`. The description of the artefact (`cei:physicalDesc`) is part of references to several artefacts confirming the existence of charter (`cei:witness`). Here, with the XML element `cei:traditioForm`, the database can describe in a very detailed way the status the version has in relationship to the engrossment. For example, *orig.* would denote that the version in question is the legally binding original. Notes like *cop.* and *ins.* denote the status of the copies. However, as classification lies with the data provider, the descriptions are highly heterogeneous and very often only determined by efficiency in the ingest process. Introducing the model developed above to control the data created and published in monasterium.net more strictly could be a path to better data quality. Additionally, the current version does not realise the description of the sequential relationship of antegraph and apograph.

Monasterium.net is different from other charter databases in that it gives the registered user the possibility to suggest changes to existing data and to create their own digital charter representations. The functionalities of monasterium.net in this respect are currently still under development, but a web-based editor (called *Edit-*

[13] monasterium.net/mom/collections/by-category#Retrodigitalisierte%20Urkundeneditionen.
[14] http://monasterium.net/mom/IT-ASFi/archive.
[15] monasterium.net/mom/DE-ABP/Urkunden/fond.
[16] www.monasterium.net/mom/RIViI/collection.
[17] www.cei.lmu.de/.
[18] github.com/icaruseu/mom-ca/blob/master/my/XRX/src/mom/app/cei/xsd/cei.xsd.

MOM3[19]) hides the syntax from the user and provides an interface which is reasonably easy to understand. A feature under development would allow the possibility to re-contextualise a charter description by linking it into a user-created collection. The user can add their own interpretations of an existing charter, re-using the image references from the original database entry. The relationship between the source and the user generated interpretation is encoded as `atom:link`. This occurs often in the Illuminated charter collection where there are extensive descriptions of the artistic decoration, usually something not provided by archives.

In the example of monasterium.net/mom/IlluminierteUrkunden/1331-05-25_Wien/charter the archives did not offer an abstract. A similar case of internal linkage is provided by RI V,1,1 n. 1730[20] which has a copy of the data in monasterium.net.[21] This copy links via the `atom:link` to the archival description[22] and additionally to the digital version of the Württembergisches Urkundenbuch (Königliches Staatsarchiv in Stuttgart 1849-1913[23]). In monasterium.net, the use of `atom:link` mark-up follows IETF-RFC4287 and allows a `type to` link to be added with the `@rel–attribute`. The IETF recommends that a controlled vocabulary is used for link relation types[24] for the values in the `@rel`-attribute, however in a charter database, it makes more sense to establish a dedicated taxonomy fitting to the model described above. Simple links such as the one to the WUB receive some semantics from the CEI-Markup. `cei:bibl` shows that it is a bibliographic reference, but it does not tell the user whether the WUB was used as antegraph, as a different description of the same charter, or if the content was partially reused.

4.2 Generalising the model

The question arises: can a general method can be found to formalise the data model in a way such that digital resources could be made aware of the versions that a charter can have and which could refer to versions of the same charter documented in several places? Some formalisations have been introduced in the description above, for example: URIs for concepts from the VID, entities and relationships from the FRBR model, XML elements from the Charters Encoding Initiative and from the Atom standard. Working with the legacy data,like that in monasterium.net, a possible solution would be to introduce controlled vocabularies for the description of links between single charter representations (`atom:link/@rel`, `cei:traditioForm`).

---

[19] github.com/icaruseu/mom-ca/wiki/How-to-Use-EditMOM3-Environment.
[20] www.regesta-imperii.de/id/1228-06-00_1_0_5_1_1_2499_1730.
[21] www.monasterium.net/mom/RIViI/1228-06-00_1_0_5_1_1_2499_1730/charter.
[22] www.monasterium.net/mom/AT-StiAStP/BlUKVariaEcclesiastica/BU_430/charter.
[23] www.wubonline.de/?wub=1129.
[24] www.iana.org/assignments/link-relations/link-relations.xhtml.

The use of persistent identifiers for single charters—for examples as URIs—lays the path to a semantic web organisation of their various relationships. In this way projects could identify the same charter when it is available in other databases and could publish lists of concordances with semantic web technologies using the `owl:sameAs` property to link between two URIs:

```
ri:1226–12–00_4_0_5_1_1_2433_1690
owl:sameAs
mom:AT–StiAStP/BlUKVariaEcclesiastica/BU_429
```

This could be extended if the projects used a common ontology of properties indicating the relationship between the charters, so a statement like the following would be possible:

```
mom:AT–StiASch/SchlierbachOCist/1411_IV_15/copy–1
dipl:authenticated_copy_of
mom:AT–StiASch/SchlierbachOCist/1411_IV_15/original
```

To support this, the results from this study of diplomatics concepts related to versioning of charters are published on GitHub as a draft in RDF (github.com/GVogeler/versioning_charters).

In addition to the relationships between different conceptualisations of a charter the ontology allows it to be stated on which level of the abstraction of a charter the data exposed is allocated, e.g.:

```
mom:AbbayeDeSaintBertin/e9944a8f–2a93–4665–a9e2–eb6c3862bf16
rdf:type
dipl:Charter_text
```

If the database can provide URIs for parts of its description, for example the transcription, the abstract and the archival reference, it could help to address this issue. With XML-data this can be achieved by assigning an ID through adding the xml:id attribute to the appropriate element and referencing it via the XPointer syntax (for example, `mom:AT–StiASchl/Urkunden/1404_II_23/#tenor` pointing to the transcription of the charter published at www.monasterium.net/mom/AT-StiASchl/Urkunden/1404_II_23/charter#tenor).

5 Conclusion

In her discussion of the possibilities of aligning charter databases according to their content, Rachel Stone concludes that it might be worth having a common data model, but concedes that the effort developing this would probably be unrealistically high. Even the VID does not cover many of the terms necessary for the classification involved in the Charlemagne project. Her argument is supported simply by the amount of possible diplomatic concepts presented in this paper under the perspective of charter versioning.

Attempts to apply the method of versioning to charters in this paper can lead to suggest the expulsion of the term *versioning* from the many core considerations. Editing, transcribing, translating, summarising, describing, transcribing, drafting, engrossing, copying, authenticating, digitising, revisioning, modifying, enhancing, and contextualising are all activities closely connected to the written artefacts documenting legal acts and all create something that can be considered a version of the *charter*. The relationship between the abstract concept of a charter with a rich intension of the term and all of those realisations can serve as the hub between them. The study suggests that it is improbable that a clear-cut definition of *charter* would serve as a starting point in the model. It seems that the conceptualisation of a charter results from a dense network of links between the things which can be easier to identify individually, for example legal acts, written artefacts, linguistic expressions, historical facts recorded by the charter, the digital representations of all of these, and even their aggregation.

This network sorts itself if the sequential feature is placed at the core of the concept of versioning. The creation of a legal fact precedes the drafting of a text, on which one or more engrossments are based. Copies, archival descriptions, and scholarly editions are created later on and can in themselves have versions, particularly in the digital realm, where copying and modifying are made easier and happen all the time. Only the sum of all those activities creates an abstract concept for charters and they all highlight different perspectives on this, including the material, the information conveyed and the linguistics.

This paper has demonstrated some approaches to serialising the data model. It seems that more data structures of the charter database would need to be exposed in a more flexible technology than the usual manner of digital representations of charters. Currently, XML and relational databases—where in both cases the data is usually displayed in HTML format—are the major forms for encoding the data structure of a digital charter representation. RDF, the semantic web data description format, is based on a graph model, which has the advantage of being able to express both data structures.

Currently no complete RDF-based model for the description of charters exists. The concepts of the VID are available as a SKOS-based knowledge base, which offers definitions of the original terminology, but it contains few hierarchical or even generic relationships. The charter projects undertaken at King's College London's department for Digital Humanities (Making of Charlemagne's Europe and People of Medieval Scotland; Broun et al.; Hammond et al.) offer a draft ontology for the legal facts[25] (Bradley and Pasin), which unfortunately contains several inconsistencies and would have to be enriched by many concepts out of the scope of the original

[25] www.michelepasin.org/ontologies/feudalism/

projects. Therefore, it is essential that a formal ontology for the description of data representing charters is created. Hopefully, diplomatics scholars will take up the challenge. Creating this ontology would contribute another important tool to aid future work on medieval and early modern charters under a digital paradigm.

Bibliography

Ansani, Michele. "Edizione digitale di fonti diplomatiche. Esperienze, modelli testuali, priorità." *Reti Medievali*, vol. 6, 2006. www.dssg.unifi.it/_RM/rivista/forum/Ansani.htm.

ArQuibanc, edited by Ignasi J. Baiges Jardí, Elena Cantarell Barella, Mireia Comas Via and Daniel Piñol Alabart, 2008/2011. www.ub.edu/arquibanc/home.html.

Barret, Sébastien, et al., editors. *Ruling the Script in the Middle Ages: Formal Aspects of Written Communication.* Utrecht Studies in Medieval Literacy 35, Turnhout, 2016.

Barbiche, Bernard. *Les actes pontificaux originaux des Archives Nationales de Paris* vol. 1, Bibliotheca Apostolica Vaticana, 1975.

Baumbach, Hendrik, and Andreas Meyer. "Lichtbildarchiv älterer Originalurkunden." *Schätze der Wissenschaft. Die Sammlungen, Museen und Archive der Philipps-Universität Marburg*, edited by Christoph Otterbeck and Joachim Schachtner, Jonas, 2014. pp. 200–7.

Bedos-Rezak, Brigitte Miriam. "Cutting Edge. The Economy of Mediality in Twelfth-Century Chirographic Writing." *Das Mittelalter*, vol. 15, 2010, pp. 134–61.

Bekiari, Chryssoula, et al., editors. *Definition of FRBROO. A Conceptual Model for Bibliographic Information in Object-Oriented Formalism.* Version 2.4, IFLA, 2015. www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf.

Berners-Lee, Tim, et al. "The Semantic Web." *Scientific American*, vol. 284, no. 5, 2001, pp. 34–43.

Bischoff, Bernhard. "Zur Frühgeschichte des mittelalterlichen Chirographum." *Archivalische Zeitschrift*, vol. 50/51, 1955, pp. 297–300.

Bischoff, Frank M. "Die Datenbank des Marburger 'Lichtbildarchivs älterer Originalurkunden bis 1250'. Systembeschreibung und Versuch einer vorläufigen statistischen Auswertung." *Fotografische Sammlungen mittelalterlicher Urkunden in Europa. Geschichte, Umfang, Aufbau und Verzeichnungsmethoden der wichtigsten Urkundenfotosammlungen, mit Beiträgen zur EDV-Erfassung von Urkunden und Fotodokumenten*, edited by Peter Rück, Thorbecke, 1989, pp. 25–70.

Böhmer, Johann Friedrich, et al., editors. *Regesta Imperii.* 98 vols., 1892–.

Borgolte, Michael. "Stiftergedenken in Kloster Dießen. Ein Beitrag zur Kritik bayerischer Traditionsbücher, mit einem Textanhang: Die Anlage der ältesten Dießener Necrologien." *Frühmittelalterliche Studien*, vol. 24, 1990, pp. 235–89.

Boyle, Leonard E. "Diplomatics." *Medieval Studies. An Introduction*, 2nd ed., edited by James M. Powell, Syracuse University Press, 1992, pp. 82–113.

Bradley, John and Michele Pasin. "Factoid-based Prosopography and Computer Ontologies. Towards an integrated approach." *Digital Scholarship in the Humanities*, vol. 30, no. 1, 2015, pp. 86–97.

Breßlau, Harry. *Handbuch der Urkundenlehre für Deutschland und Italien.* 3 vols., 2. ed., Veit, De Gruyter, 1911–1960.

Broun, Dauvit, et al. *The People of Medieval Scotland, 1093 – 1314.* 2007–2016. www.poms.ac.uk.

Buchner, Rudolf, and Franz-Josef Schmale, editors. *Ausgewählte Quellen zur deutschen Geschichte des Mittelalters*, Wissenschaftliche Buchgesellschaft, 1955–2013.

Campbell, Alistair et al., editors. *Anglo-Saxon Charters*, curr. 18 vols., Oxford University Press, 1973–2015.

Cárcel Ortí, Milagros, editor. *Vocabulaire International de la diplomatique*, 2. ed., Univ. de València, 1997.

Projet CBMA—Corpus Burgundiae Medii Aevi. Site du projet Corpus de la Bourgogne du Moyen Âge, 2004–. www.cbma-project.eu/.

Clavadetscher, Otto P. and Stefan Sonderegger, editors. *Chartularium Sangallense.* 13 vols, Thorbecke, Cavelti, 1983–2017. Available online in monasterium.net are volumes 3-11:

3. www.monasterium.net/mom/CSGIII/collection
4. www.monasterium.net/mom/CSGIV/collection
5. www.monasterium.net/mom/CSGV/collection
6. www.monasterium.net/mom/CSGVI/collection
7. www.monasterium.net/mom/CSGVII/collection
8. www.monasterium.net/mom/CSGVIII/collection
9. www.monasterium.net/mom/CSGIX/collection
10. www.monasterium.net/mom/CSGX/collection
11. www.monasterium.net/mom/CSGXI/collection.

Costamagna, Giorgio. "Dalla «charta» all'«instrumentum»." *Notariato medievale bolognese (Studi Storici sul Notariato Italiano 3,2)*, vol. 2, Consiglio nazionale del notariato, 1977, pp. 7–26.

Csendes, Peter, et al. "Register." *Lexikon des Mittelalters*, vol. 7, Artemis, 1995, pp. 581–86.

Deploige, Jeroen, et al. "Remedying the obsoloescence of digitised surveys of medieval sources. 'Narrative Sources' and 'Diplomata Belgica'. *Bulletin de la Commission Royale d'Histoire*, vol. 176, no. 1, 2010, pp. 151–66.

D'Hoop, Felix Henri. *Cartularium. Recueil des chartes du prieuré de Saint-Bertin, à Poperinghe, de ses dépendances à Bas-Warneton et à Couckelaere.* Vandecasteele-Werbrouck, 1870.

Du Cange, Domino, *Glossarium mediae et infimae latinitatis*, éd. augm., L. Favre, 1883–1887.

Enzensberger, Horst. *Beiträge zum Kanzlei- und Urkundenwesen der normannischen Herrscher Unteritaliens und Siziliens.* Laßleben, 1971.

Gervers, Michael, editor. *DEEDS (Documents of Early England Data Set).* 1975–. deeds.library.utoronto.ca/.

Gervers, et al. "The Deeds Database Of Medieval Charters: Design And Coding For The Rdbms Oracle 5". *History & Computing*, vol. 2, no. 1, 1990, pp. 1–11.

Gervers, Michael, and Michael Margolin. "The Deeds Project: Towards The Content Analysis Of English Private Charters Of The Twelfth And Thirteenth Centuries." *Le Médiéviste Et L'ordinateur*, vol. 41, 2002. lemo.irht.cnrs.fr/42/mo42_01.htm.

———. "Managing Meta-data in a Research Collection of Medieval Latin Charters." *Digitale*

Diplomatik, edited by Georg Vogeler. Böhlau, 2009, pp. 271–82.

Groß, Katharina Anna. *Visualisierte Gegenseitigkeit. Prekarien und Teilurkunden in Lotharingien im 10. und 11. Jahrhundert, (Trier, Metz, Toul, Verdun, Lüttich).* Harrasowitz, 2014.

Große, Rolf. "Rezension von 'Vocabulaire International de la Diplomatique, éd. M. Cárcel Ortí, 1994'." *Francia*, vol. 23, 1996, pp. 236–7. www.perspectivia.net/publikationen/francia/francia-retro/bsb00016300/francia-023_1-1996-00246-00246.

Hammond, Matthew, et al. "Exploring a model for the semantics of medieval legal charters." *International Journal of Humanities and Arts Computing*, vol. 11, no. 2, 2017, pp. 1–15.

Härtel, Reinhard. *Notarielle und kirchliche Urkunden im frühen und hohen Mittelalter.* Böhlau, 2011.

Hartmann, Josef, and Rudolf Engelhardt. "Zur Frage der Anwendung des Provenienzprinzips auf Urkundenbestände." *Archivmitteilungen*, vol. 14, 1964, pp. 97–107.

Heidecker, Karl, "Trois projets d'éditions informatisées d'actes aux Pays-Bas." *Le Médiéviste et l'Ordinateur*, vol. 42, 2003. lemo.irht.cnrs.fr/42/mo42_07.htm.

Hemptinne, Thérèse de, et al., editors. "Diplomata." *Belgica. Les sources diplomatiques des Pays-Bas méridionaux aux Moyen Âge.* Commission royale d'Histoire / Koninklijke Commissie voor Geschiedenis, 2015–. www.diplomata-belgica.be/.

Herold, Paul. "Ein um Form bemühtes Mißtrauen. Herstellung, Gebrauch und Verbreitung von Chirographen unter besonderer Berücksichtigung von Klosterneuburg." *Jahrbuch des Stiftes Klosterneuburg*, new series, vol. 17, 1999, pp. 153–72.

IFLA. *Functional Requirements for Bibliographic Records, by IFLA Study Group on the Functional Requirements for Bibliographic Records.* K.G. Saur Verlag, 1998, version 3.2.2: 2009. www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

Institut für Mittelalterliche Geschichte der Universität Marburg. *Lichtbildarchiv älterer deutscher Originalurkunden.* Universität Marburg, 2006–. lba.hist.uni-marburg.de/.

Isaac, Antoine, and Ed Summers, editors. "SKOS Primer." *W3C Working Group Note 18 August 2009.* www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

Jarrett, Jonathan J., and Alan Scott McKinley, editors. *Problems and possibilities of early medieval charters.* Brépols, 2013.

Johanek, Peter. "Zur rechtlichen Funktion von Traditionsnotiz, Traditionsbuch und früher Siegelurkunde." *Recht und Schrift im Mittelalter*, edited by Peter Claasen, Thorbecke, 1977, pp. 131–62.

Kölzer, Theo, et al., editors. *Regionale Urkundenbücher. Die Vorträge der 12. Tagung der Commission Internationale de Diplomatique.* Niederösterreichisches Landesarchiv, 2010.

Kölzer, Theo. "Diplomatics." *Handbook of Medieval Studies. Terms—methods—trends*, edited by Albrecht Classen, De Gruyter, 2010, pp. 405–24.

Königliches Staatsarchiv in Stuttgart, editor. *Wirttembergisches Urkundenbuch.* 11 vols, Kohlhammer, 1849–1913.

Kosto, Adam J., and Andreas Winroth, editors. *Charters, cartularies and archives. The preservation and transmission of documents in the medieval west. proceedings of a colloquium of the Commission Internationale de Diplomatique (Princeton and New York, 16 - 18 September 1999).* Papers in mediaeval studies 17, Pontificial Institute of Mediaeval Studies, 2002.

Lohse, Tillmann. "Bin ich ein Editor? Ein Selbsterfahrungskurs an der Humboldt-Universität zu Berlin." *Projektlehre im Geschichtsstudium. Verortungen, Praxisberichte und Perspektiven*, edited by Ulrike Senger, Yvonne Robel, and Thorsten Logge, Bertelsmann, 2016, pp. 182–93.

Lowe, Elias Avery. "Lay literacy in Anglo-Saxon England and the development of the chirograph." *Anglo-Saxon Manuscripts and their Heritage.* Aldershot, 1998, pp. 161–204.

Magnani, Eliana, and Marie-José Gasse-Grandjean. "CBMA – *Chartae Burgundiae Medii Aevi.* Les débuts du projet." *Bulletin du centre d'études médiévales*, vol. 9, 2005. cem.revues.org/751.

———. "CBMA – *Chartae Burgundiae Medii Aevi* I. Les fonds diplomatiques bourguignons." *Bulletin du centre d'études médiévales*, vol. 11, 2007. cem.revues.org/1064.

———. "CBMA – *Chartae Burgundiae Medii Aevi* II. Cartulaires, éditions, base de données." *Bulletin du centre d'études médiévales*, vol. 12, 2008. cem.revues.org/6962.

———. "CBMA – *Chartae Burgundiae Medii Aevi* III. Systèmes d'interrogation et recherches sur les fonds diplomatiques bourguignons." *Bulletin du centre d'études médiévales*, vol. 13, 2009, pp. 245–50. cem.revues.org/11077.

———. CBMA – *Chartae Burgundiae Medii Aevi* IV. Études, éditions, historiographie." *Bulletin du centre d'études médiévales*, vol. 14, 2010, pp. 197–207. cem.revues.org/11556.

———. "CBMA [Chartae Burgundiae Medii Aevi] – V: Actes cisterciens et prémontrés." *Bulletin du centre d'études médiévales*, vol. 15, 2011. cem.revues.org/11991.

———. "CBMA [Chartae Burgundiae Medii Aevi] – VI. Les chartes bourguignonnes sous Philologic." *Bulletin du centre d'études médiévales*, vol. 15, 2011. cem.revues.org/12047.

Mostert, Marco, and Paul Barnwell, editors. *Medieval legal process. Physical, spoken and written performance in the Middle Ages.* Brépols, 2011.

López, Martín, and María Encarnación. "La carta partida como forma de validación." *Estudis castellonencs*, vol. 6, 1994, pp. 839–55.

Mayer, Theodor. "Das deutsche Königtum und sein Wirkungsbereich." *Das Reich und Europa*, edited by Fritz Hartung, Theodor Mayer, et al., 2. ed., Koehler & Amelang, 1941, pp. 52–74

Miles, Alistair, and Sean Bechhofer, editors. "SKOS Reference." *W3C Recommendation 18 August 2009.* hrefhttp://www.w3.org/TR/2009/REC-skos-reference-20090818/www.w3.org/TR/2009/REC-skos-reference-20090818/.

Molitor, Stephan. "Das Traditionsbuch. Zur Forschungsgeschichte einer Quellengattung und zu einem Beispiel aus Südwestdeutschland." *Archiv für Diplomatik*, vol. 36, 1990, pp. 61–92.

———. "Zum Traditionsbuchwesen." *Unverrückbar für alle Zeit. Tausendjährige Schriftzeugnisse in Baden-Württemberg*, edited by Wilfried Rössling and Hansmartin Schwarzmaier, Generallandesarchiv Karlsruhe, 1992, pp. 26–29.

Muller, Samuel, et al. *Manual for the Arrangement and Description of Archives*, H. W. Wilson, 1898.

Nottingham, M., and R. Sayre, editors. "IETF-RFC4287." *The Atom Syndication Format*, IETF, December 2005. tools.ietf.org/html/rfc4287).

Parisse, Michel. "Remarques sur les chirographes et les chartes-parties antérieurs à 1120 et conservées en France." *Archiv für Diplomatik*, vol. 32, 1986, pp. 546–67.

Pergamene di Puglia Online. A cura di Carla Palma. Sopraintendenza archivistica per la puglia, 2005–. www.sapuglia.it/index.php?option=com_content&view=art-

icle&id=213&Itemid=214.

Piñol Alabart, Daniel. "Proyecto ARQUIBANC - Digitalización de archivos privados catalanes. Una herramienta para la investigación." *Digital Diplomatics. The Computer as a Tool for the Diplomatist?*, edited by Antonella Ambrosio, Sébastien Barret and Georg Vogeler, Böhlau, 2014, pp. 99–108.

Prange, Wolfgang. "Beobachtungen an den ältesten Lübecker Urkunden 1220–1230." *Lübeck 1226. Reichsfreiheit und frühe Stadt*, edited by Olaf Ahlers, Hansisches Verlagskontor Scheffler, 1976, pp. 87–96.

Rio, Alice, et al., editors. *The Making of Charlemagne's Europe.* King's College London, 2011–2015. charlemagneseurope.ac.uk/.

Roberg, Francesco, and Matthias Klipsch. "LBA online—Die Digitalisierung des Marburger Lichtbildarchivs älterer Originalurkunden." *Archivnachrichten aus Hessen*, vol. 9, no. 2, 2009, pp. 30–32.

Roland, Martin, et al. *Illuminierte Urkunden als Gesamtkunstwerk*, monasterium.net, 2015-. www.monasterium.net/mom/IlluminierteUrkunden/collection.

Rosé, Isabelle. "À propos des Chartæ Burgundiæ Medii Ævii (CBMA). Éléments de réflexion à partir d'une enquête sur la dîme en Bourgogne au Moyen Âge." *Bulletin du centre d'études médiévales*, 2008. cem.revues.org/8412.

Rübsamen, Dieter, et al., editors. *Regesta Imperii. Regestendatenbank.* Akademie der Wissenschaften und der Literatur Mainz, 2001–. regesta-imperii.de/.

Samuels, Michael, et al., editors. *Historical Thesaurus of English.* Glasgow University, 2008-. ht.ac.uk/.

Sawyer, Peter H. *Anglo-Saxon Charters. An annotated list and bibliography (Royal Historical Society / Guides and Handbooks 8).* London, Royal Historical Society, 1968.

Schulze, Ursula. *Lateinisch-Deutsche Parallelurkunden des 13. Jahrhunderts. Ein Beitrag zur Syntax der Mittelhochdeutschen Urkundensprache.* Fink, 1975.

Schwineköper, Berent. "Zur Geschichte des Provenienzprinzips." *Forschungen aus mitteldeutschen Archiven, Festschrift. H. Kretzschmar*, Rütten & Loening, 1953, pp. 48–65.

Sennis, Antonio. "Destroying Documents in the Early Middle Ages." *Problems and possibilities of early medieval charters (International Medieval Research 19)*, edited by Jonathan J. Jarrett and Alan Scott McKinley, Brépols, 2013, pp. 151–70.

Siemens, Ray, et al. "Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media." *Literary and linguistic Computing*, vol. 27, no. 4, 2012, pp. 445–61.

"Stichting Digitaal Oorkondenboek Groningen en Drenthe." *Cartago.* 2004–. www.cartago.nl/.

Stone, Rachel. "Interconnectivity of databases and charter data." *Blog 'Charlemagne's Europe,'* July 30, 2014. www.charlemagneseurope.ac.uk/blog/interconnectivity-of-databases-and-charter-data/.

Trusen, Winfried. "Chirograph und Teilurkunde im Mittelalter." *Archivalische Zeitschrift*, vol. 75, 1979, pp. 233–49.

Uhl, Bodo. "Die Bedeutung des Provenienzprinzips für Archivwissenschaft und Geschichtsforschung." *Zeitschrift für bayerische Landesgeschichte*, vol. 61, 1998, pp. 97–121.

Vasold, Gunter. "Progressive Editionen als multidimensionale Informationsräume." *Digital*

Diplomatics. The computer as a tool for diplomatist?, edited by Antonella Ambrosio, et al., Böhlau, 2014, pp. 75–88.

Vogeler, Georg. "Digital Diplomatics: The Evolution of a European Tradition or a Generic Concept?" *Studies in Historical Documents from Nepal and India*, edited by Axel Micheals et al., 2018, pp. 41–64.

———. "Von der Terminologie zur Ontologie. Das 'Vocabulare international de la diplomatique' als Ressource des Semantic Web." *Francia*, vol. 40, 2013, pp. 281–97.

———. "Digitale Urkundenbücher. Eine Bestandsaufnahme." *Archiv für Diplomatik*, vol. 56, 2010, pp. 363–92.

W3C. *Semantic Web Activities*. 2001–2013. www.w3.org/2001/sw/.

———. *Simple Knowledge Organisation System (SKOS)*. 2004–2009. www.w3.org/2004/02/skos/.

———. *Web Ontology Language (OWL)*. 2012. www.w3.org/TR/owl2-overview.

Electronic Versioning

The CMV+P Document Model, Linear Version[1]

Gioele Barabucci

Abstract

Digital documents are peculiar in that they are different things at the same time. For example, an HTML document is a series of Unicode codepoints, but also a tree-like structure, as well as a rendered image in a browser window and a series of bits stored on a physical medium. These multiple identities of digital documents not only make it difficult to discuss the evolution of documents (especially digital-born documents) in rigorous scholarly terms, it also creates practical problems for computer-based comparison tools and algorithms.

The CMV+P model addresses this problem providing a sound formalization of what a document is and how its many identities can coexist at the same time. In its linear version, described in this paper, the CMV+P model sees each document as a stack of *abstraction levels*, each composed of a) an addressable *Content*, b) a *Model* according to which the content has been recorded, and c) a set of *Variants* used for equivalence matching. The bottom of this stack is the *Physical* level, symbolizing the concrete medium that embodies the digital document. Content is moved across levels using *transformation functions*, i.e. encoding functions used to serialize (save) the document and decoding functions used to deserialize (read) it.

A practical application of the CMV+P model is its use in comparison tools, algorithms, and methods. With a clear understanding of the internal stratification of formats and models found in digital documents, comparison tools are able to focus on the most meaningful abstraction levels, providing the user with the ability to understand which comparisons are possible between two arbitrary documents.

1 Introduction

Finding differences and similarities between digital documents is fundamental for the study of digital cultural artifacts, as well as for their versioning and their preservation. With digital documents we mean both born-digital documents as well as proxy digital documents that represent other physical documents. Detecting differences between digital documents is, however, a complex task, not only because of the inherent algorithmic difficulties, but also because digital documents are stored in many different ways, using different formats and different models. For example, texts

[1] Received March 2017, published December 2019

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 153–170.

could be stored as OpenDocument files (OpenOffice), PDF files, plain-text files, Google Docs, scanned printouts, and so on.

In addition to this plethora of digital document formats, there is another complication: the *stratification* of these formats. Each document exists, at least, at two different levels: the physical level (how it has been stored on a physical carrier) and the binary level (the logical sequence of zeros and ones it is composed of). In fact, common document formats employ many more levels of abstraction on top of the binary level; for example, an XML file can be seen, at same time, as a set of XML structures, as series of characters, or a string of binary digits.

The fact that the same string of bits represents at the same time multiple views on the same document often confuses users and scholars that study documents, especially those that study how these documents have been changed over time.

This confusion extends to comparison tools as well. Tools that find and describe the differences (or the similarities) between documents are based on algorithms that focus on only one of these many possible levels: e.g., only on the binary representation or only on the XML structures. For this reason comparison tools often produce unexpected and unusable results.

Take the example of an OpenOffice document that has been converted into a Microsoft Word file: while both files contain the same content, a comparison tool will say that these two files are completely different. This is paradoxical: how can two files with the same content be completely different?

A similar "equal but different" paradox arises when we compress files. For instance, an HTML page and a copy of it that has been compressed with gzip (Gailly and Adler). We know that both files have the same content, yet comparison tools will tell us that they are 100% different. How is this possible?

The root cause of these paradoxes is the lack of a precise and formal way to describe and refer to the stratification of abstraction levels that is present in every digital document.

Without the ability to understand this stratification, comparison tools will myopically see documents at one abstraction level only, often not the one the user is interested in. The lack of such a formalization makes comparison tools also unable to compare similar pieces of information (e.g., textual content) that have been stored using different formats (e.g., ODT vs DOC).

Connected to the stratification of documents, there are not only practical issues like those just described, but also epistemological problems. Without a clear understanding of the stratification of formats and models that occurs within digital documents, it is not possible to give precise and useful definitions of key concepts such as *version*, *revision*, *difference*, *change*, or even *document*.

This paper presents the CMV+P model (Content, Model, Variants + Physical embodiment), the aim of which is to provide a rigorous, formal, precise, and actionable way

to identify and address the various levels of abstraction that exist in digital documents. Using CMV+P, humans and computers can state with precision at which level of abstraction they are performing their analysis. In the case of comparison tools, this means being able to describe which differences have been detected, on which parts of the document and at which abstraction level. Moreover, CMV+P makes it possible to meaningfully compare documents in different formats. In fact, CMV+P is a refined replacement for the document model originally designed for the Universal Delta Model (Barabucci, "Introduction"). Last, CMV+P enables scholars to reason about relations between different versions of the same document or about the evolution of documents which have changed in format or model over time.

The focus of this paper is the abstract description of the CMV+P model in its linear version. Future publications will describe more complex versions of this model for structured documents and present practical implementations.

2 How documents are written and read: an example

Before delving into the description of the CMV+P model, we should briefly discuss how digital documents (which we will refer to as simply *documents* from now on) are written (*serialized*) and read (*deserialized*). As a running example through this section and the rest of this paper, we will use a simple document that contains just the name of a fictive business: "Böh & Son."

In order to go from the concept of "a business name" to a series of bits, we will need to decide how to encode this abstract concept—or some kind of associated data structure—into a series of bits. The technical name for this process is *serialization*. As we will soon see, serializing a document consists of deciding how to describe an abstract piece of information into a less abstract piece of information. This is an iterative process: at each step we will deal with one class of details and will have to choose between multiple possibilities, all valid but with different associated trade-offs.

Media The very first choice we face is choosing which kind of media we want to use to record this name. We could make an audio recording while we pronounce the name of the business, we could draw that name (or the associated logo), or we could record it as "text."[2] In our case we will record this business name as text.

Text format Text can be stored digitally in many different ways, from *plain text* (Freed and Borenstein), where only text and no stylistic info is recorded, to more elaborate formats such as XML (Bray et al.), ODT (ISO 26300-1:2015), or PDF (ISO 32000-1:2008). To keep our example manageable we will use simple plain text.

Writing system Choosing to record the name as plain text is only the first of the choices that we have to make. Which writing system or alphabet are we going to

[2] For a thorough review of the multiple meanings of the word "text," please refer to (Sahle; Pierazzo).

use to record it? The Latin alphabet would be a common choice for people in Europe, but if we were to use that name in Japan it would be more natural to spell it using (comparable but not identical) Katakana characters. We will take the easy route and record this text using the Latin alphabet.

Character repertoire There are many ways to digitally encode a text written using the Latin alphabet in a document. The first thing to choose is a character repertoire, i.e. a standard that assigns a numeric code to each letter of the alphabet. For example, we can choose among ISO Latin-1 (ISO/IEC 8859-1:1998), Unicode (The Unicode Consortium), or CP-1252 (Microsoft Corporation). In this case we will choose Unicode and each letter will be represented by a so called Unicode *codepoint* a univocal numerical code, for instance, the letter b will be represented by the codepoint U+0062.

Character composition/decomposition In Unicode certain letters can be be encoded using various equivalent variants. In our case we have to decide how we want to encode the *ö* letter. Unicode gives us (at least) two possibilities: using the codepoint for ö (i.e., U+00F6) or using the combination of codepoints for the Latin letter o and the attached diaeresis (i.e., U+006F and U+0308). We will use the latter: separate codepoints for the letter and the diaeresis.

Byte encoding Now we must make yet another choice: which Unicode encoding should we use? In other words, how do we turn the numerical codepoints that Unicode associates with the letters into bytes? Unicode provides many possible encodings: UTF-8, UTF-16LE, UTF-16BE, UCS-32. In this example we will use UTF-8, an encoding that turns each codepoint into a group of bytes of variable length.

Byte endianness At this point, what is left to do is to turn the series of UTF-8 byte groups into a series of bytes and then into a series of bits. For this task, we will choose the so called little-endian order with 8-bit bytes. This series of bits (the so called *bitstream*) is what the computer will store on some permanent medium, for example on an hard drive.

Electron encoding However, bits are not physical entities *per se* and cannot be stored. In the case of an hard drive, bits must be stored as electric charges on a metallic plate; in the case of CDs, bits must be stored as opaque areas on the plastic substrate of the disc. In our case, we will use an hard drive whose chipset uses a simple kind of conversion from bitstream to electric states called 6b/8b (Wilamowski and Irwin). For example, the bits 110 will be stored as $-$, $+$, $+$, $+$.

Physical embodiment At this point no more choices are going to be taken. This series of electron states will be impressed by an electronic actuator on the platters of the disk. These semi-permanent alterations of the matter will be the physical carrier embodying our digital document.

Only after having gone through all these steps we can say that the business name “Böh & Son” has been serialized in an electronic document.

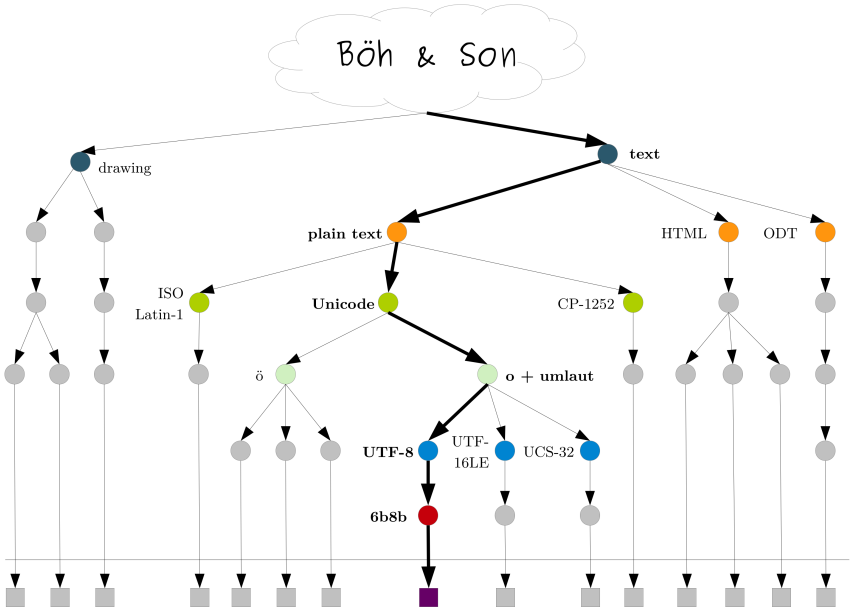


Figure 1: Tree of available choices during the creation and storage of a simple textual document. In bold the taken choices.

During the storage process we had to take many different choices. Figure 1 shows a tree of these possible choices, highlighting the choices that have been taken. In practice, however, most of these choices would not be taken by the users, but by a program, relying on clues from the user (e.g., "save the document as plain text") and following default choices hardwired in the source code by the developers (e.g., "use Unicode and UTF-8 when saving in plain text").

When an application will read this file, it will *deserialize* its content and basically undo the steps we made to write it. Pieces of information belonging to a less abstract level will be read, interpreted, and used to construct more abstract data structures, that will, in turn, be interpreted and used to construct even more abstract data structures. We see here a fundamental difference between serializing (writing) and deserializing (reading) a file. During the serialization of a document many choices are available and only few are taken. Instead, when a document is deserialized, only one set of choices can "explain" its set of physical signs (except few ambiguous cases).

We now proceed to see the details of the CMV+P model, using the example document we just created to illustrate it in practice.

3 The CMV+P model, linear version

This section describes the *linear version* of the CMV+P model, a reduced version used to describe documents whose content is not spread across different logical files.

The definitions of all the various concepts that comprise the CMV+P model will be given first. Afterwards, to illustrate how these concepts fit together in practice, the example document previously shown in section 2 will be reformulated using the CMV+P model.

3.1 Documents, abstraction levels and comparability

Definition (linear document)**.** A linear document $\mathfrak{D}$ is a potentially infinite stack of abstraction levels L_i:

$$\mathfrak{D} = (L_0, L_1, L_2, \dots)$$

For our purposes, we will limit ourselves to finite views on linear digital documents, so we will deal with documents of the form

$$(L_0, L_1, L_2, \dots, L_n)$$

where L_0 will always be the physical level and at least one of the abstraction levels will be a bitstream level.

The indexes $1, 2, \dots, n$ represent only the order in which levels are stacked in a certain document and are not meant to be compared among different stacks; in principle, the level L_3 in one document has nothing to do with the level L_3 in another document.

Our example document is thus represented by the following CMV+P document:

$$\begin{aligned} \mathfrak{D}_{ex} \quad = \quad & \Big(L_0^{physical}, L_1^{6b/8b}, L_2^{bitstream}, \\ & L_3^{UTF-8}, L_4^{Unicode}, L_5^{alphabet}, \\ & L_6^{plain-text}, L_7^{company-name} \Big) \end{aligned}$$

Definition (abstraction level)**.** An abstraction level L is a tuple composed of a set of addressable elements C, a reference model M and a set of variants V:

$$L = (C, M, V)$$

Definition (content)**.** The content of an abstraction level is a set C containing addressable elements and relations between them (e.g. order relations). The kind of elements that can be present in C and their structure are dictated by the model M.

Definition (model)**.** The model of an abstraction level is a reference M to a specification that describes what are the types of the elements in C and what are the constraints of its structure.

Definition (variants)**.** The set of variants of an abstraction level is a set V containing records of the choices, among those made available by the model M, made during the creation of C.

Definition (comparability)**.** Two abstraction levels L_a and L_b are comparable if and only if they share the same model, i.e. $M_a = M_b$.

Definition (equality between levels)**.** Two abstraction levels L_a and L_b are equal if and only if they are comparable and their contents are identical, i.e. $M_a = M_b$ and $C_a = C_b$.

Definition (equality between documents)**.** Two documents $\mathfrak{D}_a$ and $\mathfrak{D}_b$ are equal if and only if only if they contain the same number of abstraction levels and all abstraction levels of the same index are equal, i.e. $\|\mathfrak{D}_a\| = \|\mathfrak{D}_b\| = n$ and $\forall i \in \{0, \ldots, n\}\ \mathfrak{D}_a.L_i = \mathfrak{D}_b.L_i$.

Definition (equivalence between levels)**.** Two abstraction levels L_a and L_b are equivalent under the equivalence relation eqv if and only if they are comparable and all the elements that are different in C_a and C_b have associated variants v_a, v_b and these variants are equivalent under eqv, i.e. $\exists\,(c_a, c_b) \in \delta\,(C_a, C_b) \leftrightarrow \exists v_a \in V_a, \exists v_b \in V_b, eqv\,(v_a, v_b)$.

3.2 An example document in CMV+P

We can now reformulate the "Böh & Son" document described in the previous section as a stack of CMV+P abstraction levels. The stack itself is depicted in figure 2.

Let us have a look more in depth at a couple of abstraction levels, starting with the alphabet abstraction level L_5^{alphabet}. The alphabetic abstraction level L_5^{alphabet} is composed of:

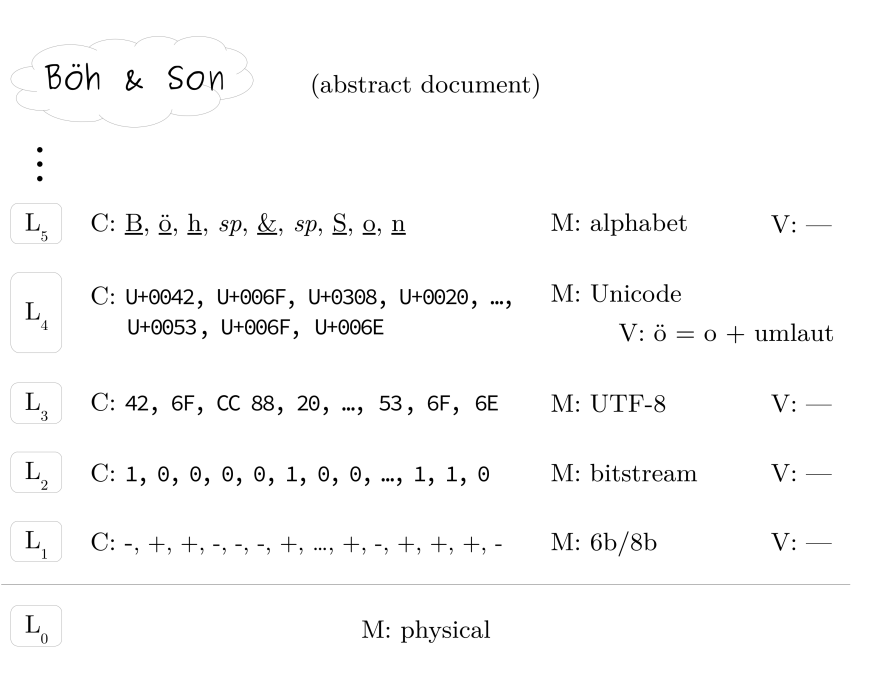


Figure 2: Abstraction levels for the "Böh & Son" plain-text document described in section 2.

- C_5, that contains an ordered list of letters (more precisely, *graphemes*), chosen among those defined in the Latin alphabet;
- M_5, a reference to the rules of the Latin alphabet and writing system (i.e. documents are composed of certain letters and punctuation signs arranged in a certain order);
- V_5, an empty set (the Latin alphabet model does not provide different but equivalent variants among which one can choose, therefore there are no choices to be made at this level of abstraction).

More formally, L_5^{alphabet} can be represented as

$$\begin{aligned} L_5^{\text{alphabet}} &= (C_5, M_5, V_5) \\ C_5 &= (\underline{\text{B}}, \underline{\ddot{\text{o}}}, \underline{\text{h}}, sp, \underline{\&}, sp, \underline{\text{S}}, \underline{\text{o}}, \underline{\text{n}}) \,, \\ M_5 &= \text{Latin alphabet}, \\ V_5 &= \{\} \end{aligned}$$

The second level we will look at is the Unicode abstraction level L_4^{Unicode}. At the Unicode level, the content is a series a so-called *codepoints*, numerical identifiers for specific glyphs (i.e. letters) as specified in the Unicode repertoire. The Unicode repertoire is a compilation of letters from many different writing systems. For instance, the codepoint allocated for the Latin letter capital A is U+0041, the codepoint for the Greek letter small beta (i.e. β) is U+03B2. For certain letters, Unicode allows for more than one codepoint, or combinations of codepoints. The Latin letter small O with diaeresis (i.e. ö) is one of these cases: it can be encoded using the single codepoint U+00D6 or the combination of codepoints U+006F and U+0308, respectively Latin letter small O and the combining diaeresis. In our example we decided to use the combining form. This will be reflected in C_4 and V_4: in C_4 two codepoints will be used to encode the letter ö; in V_4 we will record this choice.

$$
\begin{aligned}
L_4^{\text{Unicode}} &= (C_4, M_4, V_4) \\
C_4 &= \big(\text{Unicode-codepoint(U+0042)}, \\
&\quad \textbf{Unicode-codepoint}(\mathbf{U+006F}), \\
&\quad \textbf{Unicode-codepoint}(\mathbf{U+0308}), \\
&\quad \text{Unicode-codepoint(U+0020)}, \\
&\quad \ldots \\
&\quad \text{Unicode-codepoint(U+006F)}\big), \\
M_4 &= \text{Unicode, version 7.0}, \\
V_4 &= \big\{\big(\text{encode } \underline{\text{ö}} \text{ as o}^U \text{ + combining diaeresis}^U\big)\big\}
\end{aligned}
$$

3.3 Transformation functions: encoding and decoding functions

When a document is read, saved or edited, the content of all the abstraction levels that comprise a document must be kept in sync. It is thus necessary to have mechanisms that can move the content across different abstraction levels, from the document as seen via the interface by the user to the document as stored in the physical medium and vice versa. In CMV+P this mechanism is fulfilled by transformation functions.

Transformation functions are used to transform content stored according to the model of a certain abstraction level into content stored according to the model of another abstraction level. Transformation functions used during the serialization phase are called *encoding functions*, those used during the deserialization phase are called *decoding functions*. During the serialization phase, encoding functions are used to turn the content of a more abstract level into content suitable for the next less abstract level. The very last encoding function is responsible for implanting the

document into the physical carrier. During the deserialization phase, conversely, decoding functions are used to turn the serialized content into abstract data structures that the applications can work with.

Definition (transformation function)**.** A transformation function $trans$ is a function that transforms the content C_a of an abstraction level L_a (created according to model M_a and variants V_a) into the content C_b of an abstraction level L_b, created according to the model M_b and the variants V_b.

$$trans : (C_a, M_a, V_a, M_b, V_b) \to C_b$$

While the term *function* is used, it must be noted that not all transformation functions are bijective functions (i.e. complete and reversible). Some transformation functions related to the most abstract levels may not even be proper functions in a strict mathematical sense. The impact of various properties of the transformation function (e.g. bijectivity, calculability, reversibility) on the creation and interpretation of the document is out of scope for this introductory article and will be discussed in a future publication.

Transformation functions in practice

In concrete applications, the role of the transformation functions is fulfilled by various pieces of code, often embedded in shared libraries. The complexity of these function ranges from trivial to extremely intricate. For example, the encoding function from L_4^{Unicode} to $L_3^{\text{UTF-8}}$ can be written in a handful of lines of code, while the encoding function from L_5^{alphabet} to L_4^{Unicode} could consist of thousands of lines spanning a dozen libraries. A consequence of this is that, given any two abstraction levels, there exist many different concrete encoding and decoding functions between them and, in theory, an infinite number of transformation functions is possible.

Another difference between the theory and the reality is that, in theory, encoding functions and decoding functions are the mathematical inverse of each other while, in practice, the implementation of the encoding function may bear no resemblance to the implementation of the specular decoding function. Take for example a hypothetical plain-text editor software. It displays the letters that make up the text, so it must deal with L_5^{alphabet}. At the same time, the editor internally processes the textual data as Unicode codepoints at abstraction level L_4^{Unicode}. It follows that the editor must have a pair of encoding/decoding functions for these levels: an encoding function that serializes the letters of L_5^{alphabet} into the codepoints of L_4^{Unicode}, as well as a specular decoding function to deserialize L_4^{Unicode} into L_5^{alphabet}. In concrete terms, in this editor the encoding function is the code that turns input signals from the operating system

(in forms of keystrokes) into a equivalent data structures that hold sequences of Unicode codepoints; the decoding function, instead, is the code that turns the data structures that hold the Unicode codepoints into the data structures that describe the letters (or graphemes) to be displayed on the screen using an appropriate font. It is clear that these two pieces of software have little in common.

4 Using CMV+P to compare documents

Now that we have seen the basics of the model, we can move on to show how CMV+P helps in comparing documents in practice. Comparison algorithms and tools can use CMV+P to

- identify at which abstraction levels it is possible to compare two documents;
- classify which parts are identical, equivalent or different at one or more abstraction levels;
- understand which measures should be taken to compare two ostensibly incomparable documents.

To illustrate these points, this section presents a few examples of increasing complexity.

In the first two examples, the plain-text document discussed in the previous sections is compared with two slightly modified copies. Here, various kinds of differences in content and variants are analyzed. The third example shows a comparison between the same ODT file and an HTML file with similar content. This last example delves into the idea of comparing documents in different formats.

The fourth and last example deals with comparing "incomparable" documents and the associated paradox of the "equal but different" files, discussed at the beginning of this article. The purpose of this last example is to demonstrate that tools that use CMV+P can leverage their knowledge of the stacks of abstraction levels to make documents comparable by, for example, passing them through extra transformation functions.

These examples show only a few of the practical applications of the CMV+P model. Additional, more complex practical aspects of the model will be explored in future publications.

4.1 Identification of differences

Our first example deals with an elementary case of difference: a textual substitution. The first document $\mathfrak{D}_a$ contains the text "Böh & Son," the second document $\mathfrak{D}_b$ contains the text "Böh & Co." Both files are plain-text documents and have been

encoded using Unicode and UTF-8. Figure 3 shows the CMV+P stacks for these two documents.

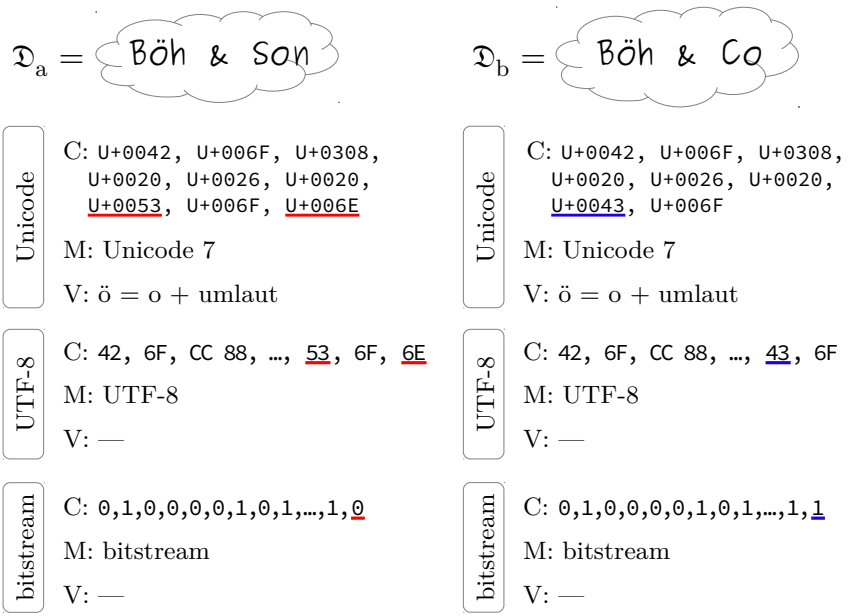


Figure 3: CMV+P stacks for two plain-text files with slightly different content.

A non CMV+P-based diff tool can focus on only one of the three abstraction levels shown in figure 3. For instance, a binary diff tool will compare only the bitstreams, while a classical text comparator will focus only the sequence of Unicode codepoints.

CMV+P allows tools to have a more holistic view of the differences. CMV+P-aware tools can provide a different set of differences for each abstraction level. For example, a tool could say: "There are three different sets of differences: at the bitstream level these bits have been changed; at the UTF-8 level these groups of bytes have been changed; at the Unicode level these codepoints have been changed." With the appropriate user interface, a single tool could provide the users with the exact kind of information they are after: the author of the text may be interested in seeing which words have changed, whereas the developer of a text-editor that is debugging a UTF-8 problem may be interested in seeing the changes expressed in terms of UTF-8 groups.

In more complex file formats, a CMV+P-aware tool has the ability to show changes at many more levels, in a clear and unambiguous way. For example, when comparing HTML files, a tool could show differences in their rendering, differences between their XML trees, differences in the textual content of various elements, or differences between XML serializations, just to name a few.

4.2 Equality, equivalence and difference

The second example illustrates how the concepts of equivalence and equality can be precisely expressed and managed thanks to the variants set V recorded in each CMV+P abstraction level.

The documents compared in this example are the plain-text document $\mathfrak{D}_a$ of the previous example and a copy of it, $\mathfrak{D}_c$, that has been serialized using the single precomposed Unicode character *ö* instead of the sequence *o + combining diaeresis*. Figure 4 shows the CMV+P stacks for $\mathfrak{D}_a$ and $\mathfrak{D}_c$.

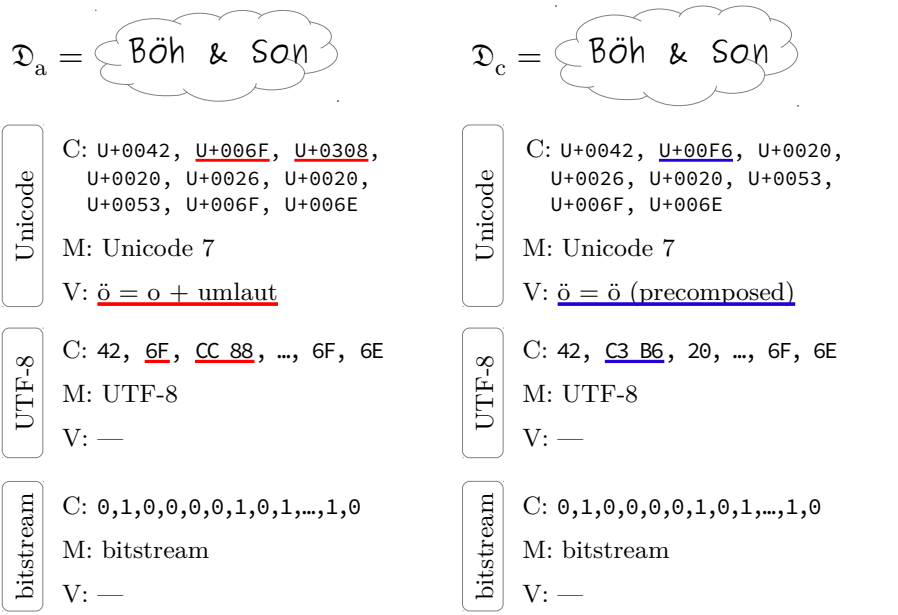


Figure 4: CMV+P stacks for two plain-text files. In $\mathfrak{D}_a$, ö is encoded with a Unicode combining character, in $\mathfrak{D}_c$ with a precomposed character.

A non CMV+P-based diff tool can either say that $\mathfrak{D}_a$ and $\mathfrak{D}_c$ are different (e.g., if it compares the bitstream of the two documents) or equal (e.g., if it prenormalizes the documents with one of the procedures suggested by the Unicode consortium (Davis and Whistler). Which of these two answers is correct depends on the needs of the user.

A CMV+P-based diff tool can, instead, provide a more complete view of the results of the comparison. It can state without ambiguity that:

- the alphabetical levels of $\mathfrak{D}_a$ and $\mathfrak{D}_b$ are identical,
- their Unicode levels are different but equivalent, and
- both their UTF-8 levels and their bitstream levels are different.

The fact that CMV+P keeps track of the set of variants used in the serialization of the documents allows formal and unambiguous definitions of what is identical and what is equivalent; c.f. the definitions in section 3. In turn, the availability of these definitions simplifies and streamlines the creation of diff tools where most of the code is format- and model-agnostic. The model-specific parts are confined to small function eqv that check the equivalence between elements that have an associated variant in the V set. Usually these functions are provided in the specifications of the model.

Performance improvements are also made possible by the existence of the variants set. Only the few variants in V need to be checked using expensive equivalence checks, the rest of the elements in C can be tested with fast equality checks.

4.3 Comparison between different formats

Allowing documents in different formats to be compared is another of the strengths of the CMV+P format. Normally, documents stored in different formats cannot be compared. For example, an HTML file cannot be compared with an ODT file, although both are basically text files with the possibility of embedding images. This example demonstrates how the use of CMV+P allows a diff tool to reason over the structure of the files being compared and to find abstraction levels that can be compared.

For this example, we will need more complex documents than the plain-text files used in the previous sections. The first document in this example is a file produced using LibreOffice in the so-called "compressed flat OpenDocument Format" (commonly referred to as "compressed flat ODT"; in the rest of this example just "ODT"). In this document, there is only a heading with the name of the business we already used in section 2: "Böh & Son." The second document is an HTML5 document with the same textual content. The CMV+P stacks of these two documents are depicted in figure 5.

Here we see that the abstraction levels of the two documents can be classified in three ways:

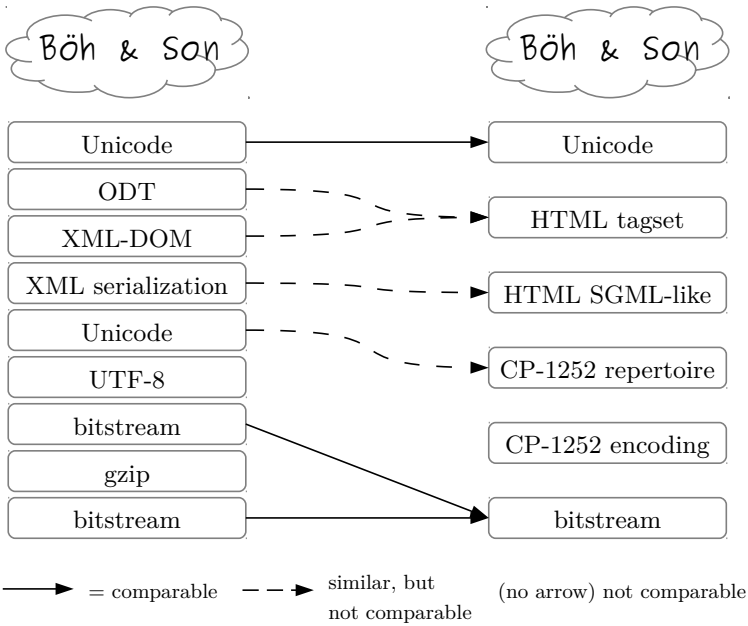


Figure 5: CMV+P stacks for an ODT and an HTML document. Only a few levels can be compared.

1. **Comparable** levels that can be compared directly because they share the same model. For instance, the Unicode levels or the bitstream levels.
2. **Similar, but not comparable** levels whose content is somehow related but that has been encoded using different models. For example, both the ODT structures and the HTML tagset have the concepts of a "heading" or a "paragraph," although they are expressed in different ways. An advanced comparison tool could compare across these similar abstraction levels if it did know about both models and had some kind of conversion function that could be used to remodel the content of these levels.
3. **Incomparable** levels whose models deal with completely different concepts, for instance gzip and HTML.

Thanks to the CMV+P model, it becomes clear at which levels comparisons can be done, where conversion functions could make two levels comparable and for which levels no comparison is possible at all.

4.4 The "equal but different" paradox solved with CMV+P

CMV+P solves the paradox enunciated in the introduction: a file and a compressed copy of it are completely different, even though they contain exactly the same content.

Let's take the plain-text document introduced in the first example (section 4.1) and compress a copy of it with gzip. The CMV+P stack of these two documents are shown in figure 6.

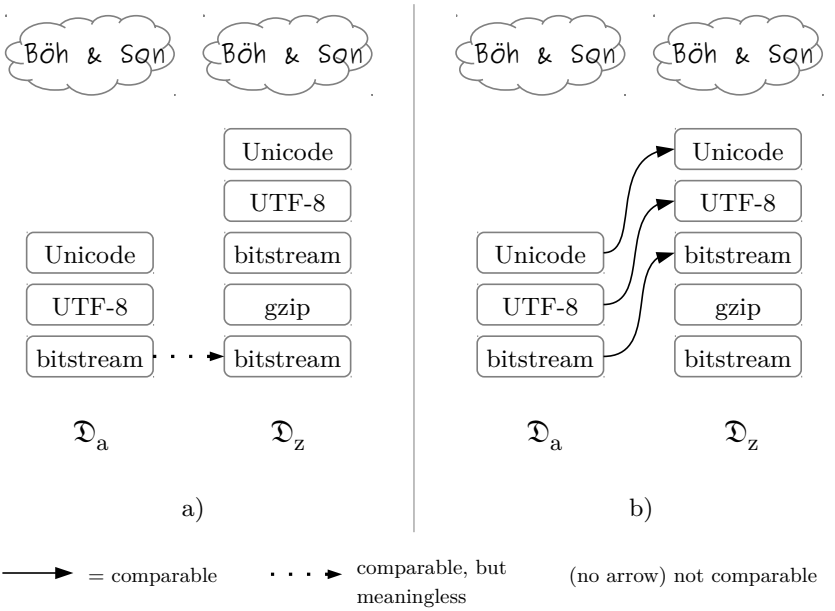


Figure 6: CMV+P stacks of $\mathfrak{D}_a$ and $\mathfrak{D}_z$. $\mathfrak{D}_a$ is a plain-text file; $\mathfrak{D}_z$ is a copy of $\mathfrak{D}_a$ that has been compressed with gzip. Subfigure a) shows how a CMV+P-aware diff tool would compare the two documents; subfigure b) shows the alignment between abstraction levels found by a CMV+P-aware tool.

All a non CMV+P-aware diff tool can do is compare the bitstream levels of $\mathfrak{D}_a$ and $\mathfrak{D}_z$. These two bitstream levels are indeed comparable, but the result of their comparison is meaningless: the series of serialized characters of $\mathfrak{D}_a$ is being compared to the quasi-random sequence of bits that is the compressed file $\mathfrak{D}_z$.

In contrast, a CMV+P-aware diff tool notices that there is a mis-alignment between the two CMV+P stacks and understands that many different meaningful comparisons are possible if the proper alignment is restored. In this particular case, the diff tool

should start the comparison process after the bitstream of $\mathfrak{D}_z$ is decompressed (or, in more precise terms, after the gzip level of $\mathfrak{D}_z$ has been deserialized using a gzip-to-uncompressed-bitstream decoding function).

A side note: aptly, in order to understand which operations are needed to make two documents comparable, CMV+P-aware diff tools (Barabucci, "diffi") must perform a sequence-alignment between the two stacks, the exact task that is at the base of almost every comparison algorithm. In other words, they have to "diff the stacks."

5 Conclusions

This paper introduced the CMV+P model (linear version) and showed that digital documents exist simultaneously at different abstraction levels.

Each abstraction level has its own peculiarities but all abstraction levels can be formally described in terms of Content, Model and Variants, together with the associated encoding and decoding functions. The final P in CMV+P reminds us that digital documents are also Physical documents, although their nature requires the use of software mediators to manipulate them.

The CMV+P model is especially useful in the context of document comparisons, in particular comparisons done with computer tools. The CMV+P model allows humans and computer tools to identify with precision

- at which abstraction levels of an electronic document a change has been detected,
- which elements of these abstraction levels have to do with this change,
- and, in general, which comparisons are possible between two electronic documents.

Bibliography

Barabucci, Gioele. "Introduction to the universal delta model." *ACM Symposium on Document Engineering DocEng '13, Florence, Italy, September 10-13, 2013*, edited by Simone Marinai and Kim Marriott, 2013, pp. 47–56, doi.acm.org/10.1145/2494266.2494284. Accessed 20. Feb. 2018.

———. "diffi: diff improved; a preview." *Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018, Halifax, NS, Canada, August 28-31, 2018*, 2018, pp. 38:1–38:4, doi.org/10.1145/3209280.3229084. Accessed 20. Sept. 2019.

Bray, Tim et al. "Extensible Markup Language (XML) 1.0 (Fifth Edition)." *Recommendation, W3C*, November 2008, www.w3.org/TR/2008/REC-xml-20081126/. Accessed 20. Feb. 2018.

Davis, Mark and Ken Whistler. "Unicode normalization forms. Standard Annex 15." *Unicode*, May 2017. unicode.org/reports/tr15/. Accessed 20. Feb. 2018.

Freed, N. and N. Borenstein. "Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types." *RFC 2046 (Draft Standard)*, November 1996, www.rfc-editor.org/rfc/rfc2046.txt.

Accessed 20. Feb. 2018. Updated by RFCs 2646, 3798, 5147, 6657, 8098.

Gailly, Jean-loup and Mark Adler. "GNU Gzip." *GNU Operating System*, 1992. www.gnu.org/software/gzip/. Accessed 20. Feb. 2018.

"ISO 26300-1:2015. Information technology – Open Document Format for Office Applications (OpenDocument) v1.2 – Part 1: OpenDocument Schema." *Standard, International Organization for Standardization*, 2015.

"ISO 32000-1:2008. Document management – Portable document format – Part 1: PDF 1.7." *Standard, International Organization for Standardization*, 2008.

"ISO/IEC 8859-1:1998. Information technology – 8-bit single-byte coded graphic character sets – Part 1: Latin alphabet No. 1." *Standard, International Organization for Standardization*, April 1998.

Microsoft Corporation. "Code page 1252 windows latin 1. Technical report, ANSI." *Microsoft Developer Network*, 1998. msdn.microsoft.com/en-us/library/cc195054.aspx. Accessed 20. Feb. 2018.

Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods.* Routledge, 2015.

Sahle, Patrick. *Digitale Editionsformen: Textbegriffe und Recodierung.* Schriften des Instituts für Dokumentologie und Editorik, vol. 9 Instituts für Dokumentologie und Editorik, Books on Demand, 2013.

The Unicode Consortium. "The Unicode Standard. Technical Report Version 6.0.0." *Unicode Consortium*, 2011. www.unicode.org/versions/Unicode6.0.0/. Accessed 20. Feb. 2018.

Wilamowski, Bogdan M. and J. David Irwin. *Industrial Communication Systems.* CRC Press, Inc., 2nd edition, 2011.

Extending Versioning in Collaborative Research

Martina Bürgermeister

Abstract

In the digital world, software development is the domain in which the management of versions, also called versioning, was first introduced and became common praxis. Versioning mechanisms usually run in the backend of web environments for data security reasons. However, what is equally important is the use of versioning as a mechanism for evidence and reference for the users. When users collaboratively create research data on web platforms, a *version history* enables others to verify and reanalyse the data. Moreover, versioning mechanisms could support scholars to make research more transparent and discursive. Following this reasoning, this paper discusses versioning as performed by version control systems and investigates how this kind of versioning can be extended to serve scholarly practice.

1 Introduction

Web-based collaborative research environments in which users create, revise, and modify contributions together, make the ephemeral status of digital resources especially evident. To counteract this shortcoming, scholars need a systematic method of organizing their resources which tracks both origin and changes. One way to do this is to automatically save all content for each change as a version with a timestamp that can be made available on request. Equally important in the research process is the management of these versions, including mechanisms to persistently refer to and access individual versions. Version management enables others to verify, reanalyse, and reference the data generated with each edit, increasing the transparancy of reaserach processes.

In the digital world, software development was the domain where the management of versions, or versioning, was first introduced. Designed as a support tool during the software development process, in a versioning system all changes are logged and assigned a unique number or name in relation to the current developmental state, so that they are always ready to be recovered at a later moment. Similar versioning mechanisms run in the backend of web environments for data security reasons. Versioning management provided by Version Control Systems (VCSs) were first developed for the software industry to make production more efficient, but they have a benefit that clearly goes beyond their original use:

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 171–190.

> […] not only does using a VCS solve many common problems when writing code, it can also improve the scientific process. By tracking your code development with a VCS and hosting it online, you are performing science that is more transparent, reproducible, and open to collaboration (Blischak et al.).

How can this benefit of VCSs be applied to research environments in the humanities? How can these systems support collaborative research? And are they good enough? Firstly, this paper will define collaborative environments in research and highlight some examples from the field of Digital Humanities (DH). Secondly, the paper will take a closer look at VCSs in general and at *Wikipedia*'s adoption of a software development approach to versioning for the purposes of collaborative editing. Starting from this example it will be shown that versioning systems can be extended for research purposes and how they could be implemented. Finally, the paper will juxtapose three ontologies which are frequently used in DH and in the cultural heritage sector (FRBRoo, OAI-ORE and PROV-O), and how these ontologies approach versioning-like concepts. However, a finished implementation will not be provided in this paper.

2 Web-based collaborative research in DH

The importance of dealing with changes in digitally published web content becomes apparent especially within wiki-like research environments. For the purposes of this paper, Carusi's and Reimer's definition for such environments will be used:

> To summarise, we found that the term used was not important, though the understandings associated with the terms *VRE*, *Collaboratory* and *Gateway* are converging on a set of characteristic features: an electronic web-based environment for a) access to data, tools, resources; b) co-operation or collaboration with other researchers at the same or different institutions; c) cooperation at the intra- and inter-institutional levels; or d) preserving or taking care of data and other outputs. Not all of these environments serve all of these functions, but they generally serve two or more (15).

The focus of this paper is on web-based environments where content is collaboratively generated and should be preserved and be openly available. Henceforth the term *collaboratory* will be used for such environments. Virtually anyone could be a potential collaboratory contributor: researchers, students, or interested laypersons.

The following examples of collaboratories are taken from the DH field and include, to a certain extent, a versioning strategy.

One of the earliest collaboratories in this context was *Suda On Line*,[1] an attempt to collaboratively translate the *Suda*, a Byzantine Greek historical encyclopaedia. *Suda On Line* started at the end of the 1990s and the last translation was published in 2014. At that time, the platform had about 200 contributors from more than 20 countries (Suda On Line). Registered users could edit and translate passages of the *Suda*. Administrators revised the translation or any changes made and published it on the platform. Despite not being able to retrieve older versions, a basic type of version history is presented: it is possible to know who made the entry, who vetted the transcription, and when it was written.

Another example is *Papyri.info*,[2] which was launched in 2006. *Papyri.info* aggregates papyrological material from different databases and makes them available and describable. Users of *Papyri.info* can browse through the distributed resources and, when registered, add new descriptions or change existing ones. There is also a peer-review process for the edits. The whole project is managed using *Git*:[3] all edits are thus recorded, versioned, and recoverable. *Papyri.info* provides a full editorial history by linking to the software development platform project site.

Annotated Books Online (ABO)[4] is a virtual research environment for scholars and students. It was launched in 2013 and is part of the research project *A Collaboratory for the Study of Reading and the Circulation of Ideas in Early Modern Europe*. *ABO* contributes to the study of marginalia and enables the tracing of reading practices and the use of books (Visser 67–69). Registered users of *ABO* can also transcribe and translate the marginalia. All transcriptions and translations are supervised by the project administrators, who guarantee the quality of the contributions. Each contributor can re-edit an already edited area. There is a basic edit history, which displays who created the entry and when the record was last modified.

The last collaboratory[5] considered here is *Monasterium.net*.[6] It was founded in 2001 and it is Europe's largest collaborative archive for charters from the middle ages and the early modern period. *Monasterium.net* allows diplomatists, archivists, and even interested laypersons to share their research with the general public. *Monasterium.net* does not provide any version history and former versions can no longer be displayed, but it does enable different versions to coexist as interpretations of the archival

1 www.stoa.org/sol.
2 www.papyri.info.
3 Git is one of the most used version control systems at the moment. The repository of the papyri.info editing framework was initiated in 2008 under: github.com/papyri/sosol/graphs/contributors.
4 www.annotatedbooksonline.com
5 For more on collaborative projects in DH see, for example, Deegan, Neuroth.
6 www.monasterium.net.

objects. It is possible that an archivist with a special interest writes one abstract and a diplomatist or philologist contributes another separate view of the same charter.[7] This concept of "parallel versions"[8] will be relevant later on in this paper.

3 Versioning in software development

In software development, versioning is implemented by version control systems (VCSs), first deployed in the 1970s.[9] Complex software programs consist of millions of lines of code, which are intended to function smoothly together. Independent of programming style, software development is an iterative process: the developer writes the code, tests it, rewrites and enhances it, and then the cycle starts again with the testing phase. Version control software tracks the iterations or changes made to the files at each step of this iterative process providing access to each version of the file. Each version has a timestamp and an author. At any time, it is possible to identify who changed what and when. All changes, accidental or not, can be reverted, and old statuses of files can be recovered. In a collaborative software development setting, people have the possibility to work simultaneously on the same code file(s). In this case a VCS coordinates possible conflicts within the code. The rationale of these systems is the avoidance of data loss (Baerisch 1–9).

Roughly grouped, there are two system architectures used for VCSs. On the one hand, there is the so-called centralized approach, which has been the standard for version control for many years. A code repository on a single server contains all the versioned files and clients receive the code from that central system. A well-known representative is *Apache Subversion*.[10] On the other hand, there are distributed VCSs, such as *Git*, *Mercurial*, or *Bazaar*.[11] Distributed systems allow the clients to have a full copy of all files in the repository on their local machines, with the advantage that in the case of a server crash, every client can restore the data (Chacon and Straub).

All VCSs share similar concepts and common functionalities. One of these is that they provide repositories on one or more servers. These repositories contain all code files of a software project, which are administrated and versioned by the VCSs. Every user can have an up-to-date working copy of the dataset, which is the current version. The server administrates the repository and coordinates the versions from the clients and keeps all the versioning information.

Anyone who wants to work on the files must first check them out from the project's code repository. This means the user gets a copy of the files and can choose if she

---

7 The importance of this feature is debated by Georg Vogeler: Vogeler 74.

8 "Parallel versions" in markup are discussed, e.g. by Buzzetti or by Vasold, as "Informationsräume."

9 See i.g. Source Code Control System: sccs.sourceforge.net.

10 subversion.apache.org.

11 github.com; www.mercurial-scm.org; bazaar.canonical.com/en/.

wants the latest or an earlier version of the code. In any case, the user can start to edit the local copy. If the changes are to be brought under version control again, the user sends a *commit* request to the VCS. The new code is then compared with the checked-out-code and the differences are saved, or like with Git, a snapshot of the files is recorded with a reference to its preceding version. This new code can be fetched from the repository by all other users. Hence, all users stay in sync with their data and get the latest updates. The update process is similar to the commit: the local project copy is compared with the current repository project and differences are merged into the local version.

When different users are simultaneously working on the same file, there is the potential (even the inevitability) that their changes will come into conflict.[12] In the case of multiple users working on the same lines of code and with concurrent changes being made on the same file as a result, the VCS does not merge these changes automatically on update. It reports the conflicting parts of the code and the user is asked how to integrate the relevant sections of code in order to solve the conflict. The changes to the local file copy will not be reflected in the repository until the user merges the conflicting parts in the file (Baerisch 9–15).

Branching in the development process supports experimentation with new ideas, such as when testing different feature implementations. The advantage of a *branch* is that changes in one branch do not affect other branches in the repository. A branch in this context means that at a certain point in the development two parallel versions of code are developed separately. In the introduction to the VCS called *Source Code Control System*, Eric Allman in 1980 (Nyman 73) notes "Creating a branch 'forks off' a version of the program." By this process a copy of a program is made, which the current software development community a *fork*, though the differences between branches and forks are not so clear in the literature (compare Nyman). A fork, according to Robles and González-Barahona (3), happens when "a part of a development community (or a third party not related to the project) starts a completely independent line of development based on the source code basis of the project." Furthermore, they argue that for something to be called a fork, there should be the following conditions: a new project name; a branch of the software; a parallel infrastructure; and a new developer community (Robles and González-Barahona 3). However, there are also more general definitions, as Nyman found in his studies and interviews:

> Indeed, in addition to being used by some as a synonym with branch, fork may also now be seen used among developers to indicate to reusing [sic] existing code in the creation of a program that may target a significantly

[12] Although most modern VCS support collaborative editing, in more restrictive types of VCSs, this is not possible because only one user has editing rights to a file at the same time.

different user group than the original developed by hackers with no affiliation with the original project (Nyman 132).

To give a better understanding of versions, branches and forks in VCSs, the following figures summarise the main concepts and their consequent structures: Versions saved in a VCS form a line of versions (fig. 1). Versions occur in a sequence of time and are based on past versions.

Figure 1: Line of versions. In grey the most recent version.

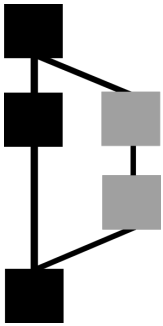


Figure 2: Temporary Branch in grey merged with the main branch.

A branch shapes the sequence of versions as a tree. Versions are developed in parallel (fig. 2). There is no definitive way to deal with branches (compare Nyman) but in general these are of temporary nature – either merged with the main branch (fig. 2) or deleted.

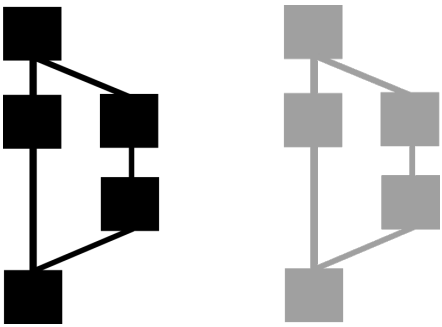


Figure 3: Original in black and fork in grey.

Forking can have different meanings within VCSs. In the *Git* model, forking belongs to the normal developer workflow when contributing to other open software projects (Chacon and Straub 123). A fork is thus a parallel project. The code base is copied with its history and developed in different ways independently from each other (fig. 3).

Finally, version control is not only used in software development, but also in the context of web content versioning. For example, *Wikipedia* uses a VCS to manage all user contributions and to display this information to them in the form of a history page.

4 Versioning in Wikipedia

Wikipedia is one of the most visited web sites and popular collaborative editing platforms, which has been called "history's biggest experiment in collaborative knowledge" (Poe). *Wikipedia* entries are created and updated every few seconds[13], clearly distinguishing it from paper-based encyclopaedias. Additionally, it operates in more than 280 languages. Most of the lemmas explained on the platform exist in different languages not just as translations, but with its own original content for each language, no matter in what version. Consequently, the content may differ significantly: in some languages a concept is described in a well-researched way, whereas others have more rudimentary explanations, which reflect the *Wiki* philosophy of, and approach to, free content: "Wikipedia has no firm rules. Wikipedia has policies and guidelines, but they are not carved in stone" (Wikipedia contributors, "Five pillars").

Wikipedia is largely written by interested laypeople, especially if the topic of the article is of general interest such as articles on pop-stars, movies, etc. There are no restrictions on contributors or content, and this fact can lead to a lively discussion process, and even to *edit wars*, when differences in perspective cannot be resolved.[14] *Wikipedia* is aware of this problem and therefore reserves the right to close articles for editing. Before doing so, those involved are called upon to observe *Wikipedia etiquette*. *Wikipedia* is aware that articles are sometimes of poor quality (Wikipedia contributors, "Reliability") and may contain false information, present only one single point of view, and the coverage of subjects might be heavily unbalanced. This happens when editors follow a special interest and ignore the requirement for an impartial and comprehensive perspective. However, *Wikipedia* believes in "its self-healing effects," and interprets the aphorism of "Given enough eyeballs all bugs are shallow" (Raymond 30) as representing the way in which all content will improve over time

[13] On the 26^{th} March 2017, the English *Wikipedia* had 5,368,820 content articles and 41,786,208 pages in total that were generated by 881,544,043 edits. There are 30,550,216 registered users with 1,268 administrators (Wikipedia contributors, "About").

[14] On edit wars, see: Kallass 305–309.

given enough edits (Wikipedia contributors, “About”). Nevertheless, it is frequently necessary to reverse malicious edits, and for that purpose a VCS saves versions of all edits made to an article.

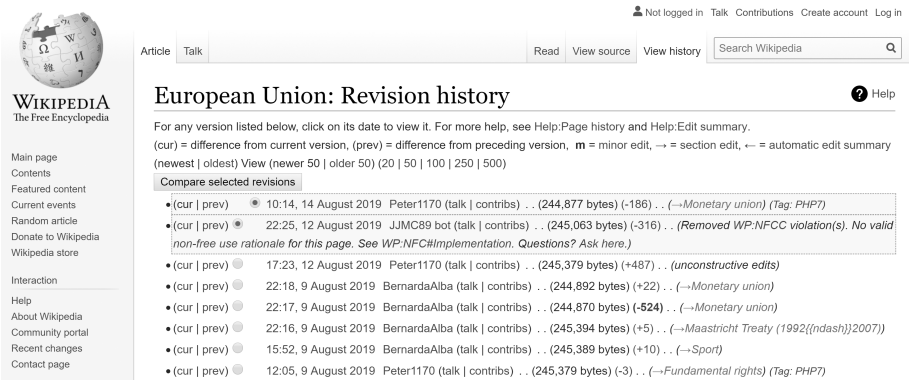


Figure 4: Revision history page of Wikipedia.

All text versions of a specific *Wikipedia* page are accessible to the users via a history. The revision history page (fig. 4) lists all changes of any editable *Wikipedia* page. The edits are listed from newest to oldest, and for each edit the following details are listed: time and date of the edit, username of the editor or IP address for not registered users, the size of the file, the sum of deletions and insertions and optionally an edit comment. In the case of long edit histories, the user can navigate the versions by year and month. There are links to the most recent and oldest edits and the previous and next page in the history. Each listed version is shown in comparison to the current version by clicking “(cur)”, while by clicking the “(prev)” link next to each version, the previous version is shown on a separate page. Any two versions on the page can be compared by clicking the radio buttons in the lines of the selected versions. *Wikipedia* distinguishes between minor and major changes: a bold “m” in the revision history signals that only minor changes have been made to the preceeding version. *Wikipedia* defines minor and major edits as follows:

> A check to the minor edit box signifies that only superficial differences exist between the current and previous versions. Examples include typographical corrections, formatting and presentational changes, and rearrangements of text without modification of its content. A minor edit is one that the editor believes requires no review and could never be the subject of a dispute. An edit of this kind is marked in its page’s revision history with a lower case, bolded “m” character (**m**).

> By contrast, a major edit is one that should be reviewed for its acceptability by all the editors concerned. Any change that affects the meaning of an article is not minor, even if it concerns a single word; for example, the addition or removal of "not" is not a minor edit (Wikipedia contributors, "Minor edit").

These categories allow for a weighting between more and less substantial changes in a version history. The most interesting changes are those emerging from a debate in which an editor succeeds in imposing her will. This usually happens because it is the strategy of *Wikipedia* to develop one article per lemma and conflicting parts arising from different percptions have to be consented to and merged. It is the same with conflicting code in software development unless you fork. This is not the case in research praxis, where a multiplicity of opinions and perspectives is important.

5 Versioning scholarly positions

As mentioned above, the collaboratively generated content can be overwritten, corrected, or continued; however, in all instances just the most recent version is displayed. This happens when the versioning strategy is realized as a one-to-one relationship of versions. The version represents one possible intellectual position. But, what if we want to represent different research positions or interpretations of the same edited object? Should there always be a privilege for the most recent version? In a scholarly discourse this is not acceptable. However, if we want to use VCSs to make scholarly positions transparent and documented, it is necessary to adopt a versioning strategy, which enables multiple perspectives as well as dissent. Meister defines key requirements for a collaborative approach to textual markup which can be effective within collaboratories as well. In a collaborative setting, it should be possible to relate one source object with more than one describing instance. Meister states:

> Collaborative markup must be based on a one-to-many relationship model in which one object text is related to *n* associated markup instances [...] every markup instance should be preserved as unique data set, with a pointing relation to the source document. [...] All original source documents and all the markup instances related to these must be handled by an integrated document management system [...] (120)

VCS could handle document management. As mentioned above, VCSs provide the ability to develop and manage software code in parallel. Especially forking, in the sense of spin-off and new development, offers new possibilities for the scholarly discourse. The "right to fork" (Nyman 1), which is so important for the open source software community, can be adopted for scholarly purposes. With a fork, you have the possibility to develop in a new direction something independent from the predecessor

or origin. The crucial difference from just copying the research object is that, in the case of a fork, the reference to the origin that is being copied remains. This means that VCSs could easily manage different views on research objects.

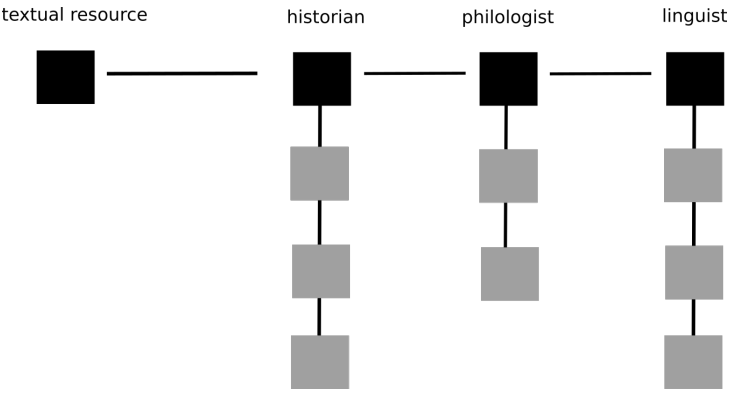


Figure 5: Forks as parallel positions in black, individual versions in grey.

In this way, as illustrated in figure 5, a textual resource can be forked and as such analysed by, e.g., a historian. A philologist, on the other hand, could set up a fork with his or her interest and questions, and similarly other researchers with different interests and points of view could do the same. This means that by forking through a VCS, scientific discourse is trackable, thus guaranteeing the plurality of interpretations.

In a VCS versions are administrated entities without further semantic meaning. Each change is a new version in the project under version control. Users can compare versions and see what has been changed; afterwards, they may evaluate and interpret the changes. In general usage, as a term, version is versatile and not very precisely defined (see Nury in this book). In some contexts correction of spelling may not be considered a reason to form a new version, but for others they are—and VCSs agree with the latter, since every saved change creates a new version.

The concept of a fork (respectively branch) helps to distinguish semantically and technically from versions. Forks specify parallel versions. Parallel versions have the same original versions, but do not overwrite each other every time they are changed. However, similar to versions, the meaning of each fork and its purpose can vary greatly and we do not know its specific purpose. Only in the comparison between the fork and its origin or with other forks does the meaning become clear. Would it not be desirable to determine the type of change, to know the reason for a parallel development? In concrete terms, would it not be interesting for the scientific discourse

to relate the different versions in order to make the significance of each version and of their contribution to the full collaborative process clear?

An approach in software development that takes the semantics of changes into account is *Semantic Versioning*, which suggests a version notation of software releases in order to more easily determine compatibilities with different software packages. Tom Preston-Werner's prescription is as follows:

> Given a version number MAJOR.MINOR.PATCH, increment the: MAJOR version when you make incompatible API changes, MINOR version when you add functionality in a backwards-compatible manner, and PATCH version when you make backwards-compatible bug fixes.

Following his criteria, the notation is meaningful for all developers and users. He differentiates three types of versions. With the classical VCS, we can differentiate two types: a version and a parallel version. However, if we want versions to be meaningful, they must be differentiated in relation to other versions, such as: "this is an alternative point of view," "this is a contraposition," "this is an addition," or "just revision and not really a new version."

In the following, I will present well-known data models in the field of DH, which can extend the traditional versioning model. They promise to document dynamic objects and their changes over time. Each of the three examples comes from a different domain, and each has its own concept and way of dealing with change, as well as processes that produce versions: FRBRoo, OAI-ORE and PROV-O.

6 Three examples for modelling versions

6.1 Versioning with FRBRoo

The knowledge of libraries and museums is represented in the FRBRoo ontology, which was published in 2009 as an extension of CIDOC CRM[15] and FRBR.[16] FRBRoo is a very large ontology; however in this section, just a few aspects are considered because of their relevance to the modelling of versions. Let us first look at the concepts of *Work* and *Expression*.[17] The notion of *Work* is described as:

> A distinct intellectual or artistic creation. A *work* is an abstract entity; there is no single material object one can point to as the *work*. We recognize the *work* through individual realizations or *expressions* of the *work*, but the *work*

[15] International Council for Museums – International Committee on Documentation. Ifla.com

[16] Functional Requirements for Bibliographic Records. Ifla.com.

[17] In FRBR and FRBRoo Manifestation and Item do not include web resources. An adaption for digital documents is suggested by Albertsen and Van Nuys.

> itself exists only in the commonality of content between and among the various *expressions* of the *work* (IFLA 54).

Expression is already mentioned in the above definition as a realisation of a *Work*. More fully:

> Inasmuch as the form of *expression* is an inherent characteristic of the *expression*, any change in form (e.g., from alpha-numeric notation to spoken word) results in a new *expression*. Similarly, changes in the intellectual conventions or instruments that are employed to express a *work* (e.g., translation from one language to another) result in the production of a new *expression*. If a text is revised or modified, the resulting *expression* is considered to be a new *expression*. Minor changes, such as corrections of spelling and punctuation, etc., may be considered as variations within the same *expression* (IFLA 55).

This definition comes from a real world problem for librarians: obviously, there are differences from the world of software developers and the use of VCSs. What in a VCS is a version, in FRBRoo is either a variation within the same expression or a new expression. Not every change generates a new version as expression: minor changes would not have an impact on the expression.

In FRBRoo, dynamic aspects of objects can be represented by *events*. Events bear the chance to identify similarities of, or differences between, objects. By describing the event of e.g. creation, which is situated in time and space and involves actors, FRBRoo expression becomes a comparable or exchangeable information (fig. 6). There are concepts for events that deal especially with the creation, publication and production processes.[18]

Applied to the process of versioning, all versions are expressions and would be generated by events, because it is the activity *commit* that brings a version into existence. Furthermore, the version that is generated by the creation event could be modelled with a timestamp and a responsible person. Then in a collaborative research setting new versions could be created by modifying the former version. In this case, a type like "Revision" or "Adaption" can be added. In order to describe the relationships between versions' properties, such as "consists of," "influenced by," "has modified," "is composed of," "refers to," "is logical successor of," "is derivative of," or "is based on" could be used. FRBRoo and CIDOC CRM allow abundant possibilities to describe version relationships between cultural heritage objects, especially because both provide more than 230 different relationships between entities.

---

[18] Besides these events, there are terms for "Recording Event," "Performance," and "Reproduction Event" (IFLA).

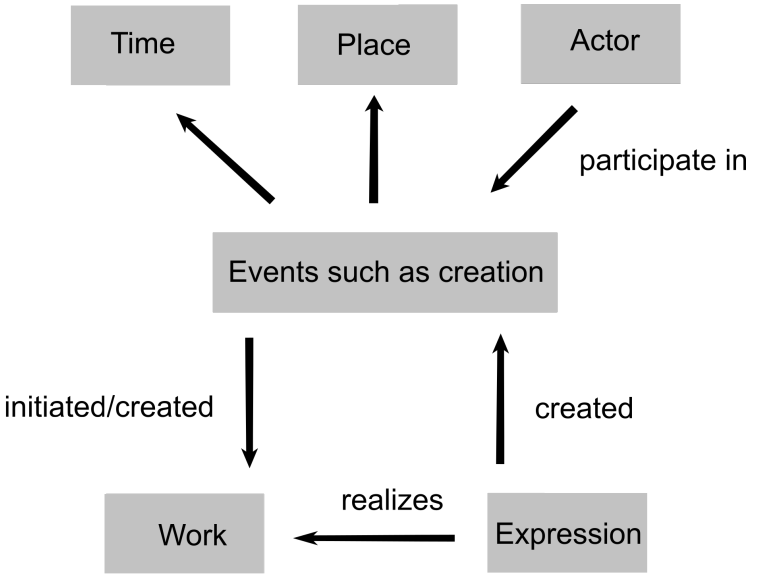


Figure 6: Temporal entities modelling intellectual processes (simplified).

6.2 Versioning with the Open Archive Initiative

The *Open Archive Initiative* (OAI) has its roots in the development of e-print repositories, so-called archives. Version 1.0 of the *Object Reuse and Exchange (ORE)* data model specification was released in 2008 and no further version has been released yet.

The strength of the *OAI-ORE* is in particular in describing and exchanging web resources. It aims to provide concepts to describe their constituents and boundaries. Furthermore, it is possible to easily compound information objects to a logical whole (Lagoze and Van de Sompel, "Compound"). Usually web resources are *compound information objects*. They can be located on a single web server or be distributed anywhere. However, these resources are somehow related to each other forming a logical entity and OAI-ORE is used to encapsulate and document their relationships. An example can give an idea of what is meant by this: an entry on a collaboratory can be described as a combination of all involved software, interlinked HTML pages, facsimiles and text documents as well as all versions of software, HTML pages, facsimiles and texts. All these resources can be convened through one information object, whose components and relationships are described together.

Compared to *FRBRoo* or *CIDOC CRM*, the *OAI-ORE* abstract data model is lightweight. The model differentiates four classes, representing the core entities of interest:

firstly, the class *Aggregation*, which is a pure conceptual construct that aggregates the set of interesting resources. Van de Sompel and Lagoze call it a containment node ("Archive"), i.e. a fictive, empty node with a HTTP URI that binds the resources to the compound object and as such defines its boundaries as well (Open Archive Initiative).

An Aggregation is a compound of resources. One of these resources is an instance of the class *Aggregated Resource*. All Aggregated Resources, in turn, constitute the Aggregation. The Aggregated Resources can differ in media type and format. Each Aggregated Resource has its own URI. All resources belonging to one compound information object are listed in a *Resource Map* (Open Archive Initiative).

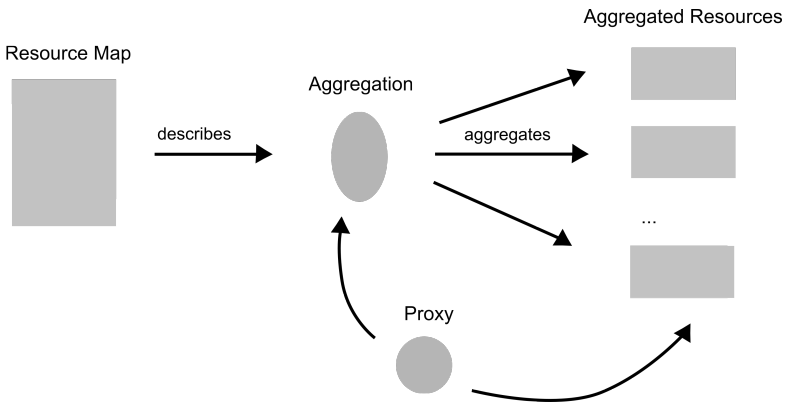


Figure 7: Illustrating four classes of the OAI-ORE model.

The *Resource Map class* contains all the information for machine consumption. It is an encoded description of the compound information object, which has a separate URI. The Resource Map lists the Aggregation and all Aggregated Resources, relates the single resources to each other, and assigns properties. As part of the obligatory description, it contains metadata of the author as well as a date-timestamp. The Resource Map provides the description of exactly one Aggregation, which forms one of various possible representations of the combined resources (Open Archive Initiative).

The *Proxy* class provides a method for ordering the resources within compound objects. The Aggregated Resources do not have an inherent order for serializing the data: in the *OAI-ORE* model this is done with the help of *Proxies*. The relationships between Proxies define the order of the resources and the Aggregated Resources, which they represent, stay neutral, without further context information. A Proxy has its own URI name, which is listed together with all the other information in the Resource Map (Open Archive Initiative).

In a collaboratory, research tracked with OAI-ORE would enable the sharing and reuse of resources. Every user creates her own set of resources: for example, a source (images) and corresponding research results (text documents). Different versions of text, like intermediate results or partial results, and images would be separately addressable resources, which can be sorted chronologically by proxies. When another collaborator continues working on a version (resource), it is aggregated and is therefore part of another set. When work continues on this aggregated version, a new resource is created - represented by a proxy and linked to the previous version. This means that any number of descriptions can be created in a collaborative research environment. References to other resources (versions) can be created. Even a whole set can be aggregated again as a resource and thus become an object of further consideration. This fosters a free give and take that "provide(s) the foundation for advanced scholarly communication systems that allow the flexible reuse and refactoring of rich scholarly artifacts and their components"(ORE Specification). When it comes to the exact description of relationships, OAI-ORE does not offer many possibilities. However, the standard is open to integration with other vocabularies that provide useful terms in their domains (Open Archive Initiative), a process that would be necessary for an adequate description of the versions. The *OAI-ORE* data model recommends the use of *Dublin Core Elements*, *Dublin Core Terms*, *Friend of a Friend* terms, *RDF* terms, and *RDF Schema* terms, as well as terms from *PROV-O*, which is presented in the next section.

6.3 Versioning with the Provenance Ontology

The *Provenance Ontology* (PROV-O) is a domain-agnostic ontology, recommended by the *World Wide Web Consortium* (W3C)[19] in 2013. Provenance is defined as "a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing" (W3C). In any case, it provides a better understanding of contexts. Provenance information intends to increase the trustworthiness, reliability and quality of the described resources (W3C). The main purpose of the *PROV* is to provide a standardised way for representing and exchanging domain-independent provenance information. In its core, the *PROV* data model is about *entities*, *activities*, and *agents*, who are interconnected by relationships (see fig. 8).

Entities are all things of interest about which we ask provenance information. These can be physical items such as a printed book or any other artefact, but also abstract ideas and digital objects as web pages and files. The definition is: "An entity is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities

[19] Founded in 1994, the W3C develops standards for the World Wide Web: www.w3.org.

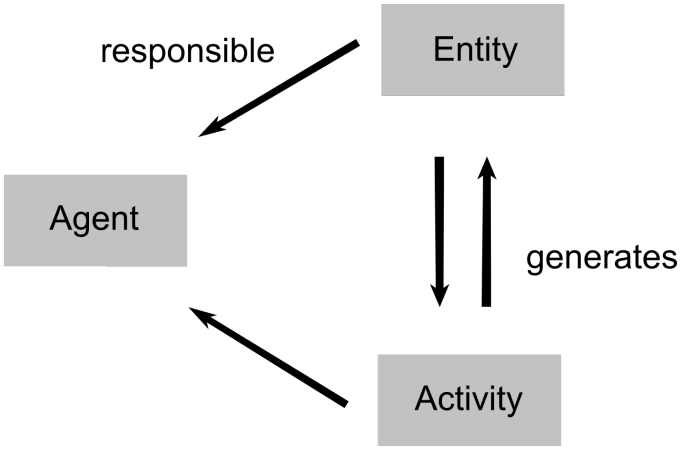


Figure 8: Core concepts of PROV-DM.

may be real or imaginary" (W3C). Instead, an *Activity* is something that occurs to entities, interacts with entities over a period of time or produces new entities. "It may include consuming, processing, transforming, modifying, relocating, or generating entities" (W3C).

An entity which is based on a pre-existing object is called a *Derivation*. A special case of derivation is the concept of *Revision*. A Revision is an entity that implicitly contains substantial content of the original entity: it "is a derivation for which the resulting entity is a revised version of some original" (W3C).

Provenance is described from different points of view, with different foci on agents or entities and their activities. Hence, the ontology considers concepts that are about distinguishing the same thing under different conditions and perspectives. When there is more than one provenance description for the same thing, these description entities can be related through the concepts *Specialization* and *Alternate*.

Right after the last release of the PROV-DM in 2013, De Nies et al. wrote a web service to convert processes in a VCS into PROV. In order to make provenance information within VCS interoperable and exchangeable, Arndt et al. continue and extend this approach and suggest a semantic representation of a Git commit, which as RDF-graph can be queried via a public API. They mapped a commit as an activity, associated with an author and a committer and with a start and an end time. Since they were working to version RDF-graphs, their entities were named graphs in a file linked to the commit via *wasGeneratedBy* and which attributed its origin as *specializationOf*. They further enriched the functionalities of Git with metadata; their

approach, however, does not describe the semantic meaning of the individual versions. PROV provides the concepts *Revision*, *Derivation* or *Alternate* exactly for this purpose, to avoid unspecific versions.

7 Conclusion

Over the last 15 years, collaboratories have brought together the efforts of many contributors to collaboratively conduct and share research. However, in order to make collaboratively created research output and its creation process transparent, all edits have to be managed and made citable and accessible. A minimal solution would be that collaboratories provide a version history to their users. Many of the platforms already have this kind of information because their databases are under version control and it is important to pass this information on to the users in order to verify and replicate the various contributions. It must be possible to document the issues we communicate on with a standardized approach and in such a way that they are made accessible for all.

A challenge still unresolved is the need to work together while allowing contributors to follow different interests and investigate different research questions. If we wish to establish a scholarly discourse in collaboratories, we have to create the space for multiple opinions. Yet we cannot apply *Wikipedia*'s principle of *Linus' law* i.e. all viewpoints must be merged in order to reach a consensus. I would argue that the *Wikipedia* principle is not great for research platforms and a better approach should be in concordance with Raymond's statement: *Given enough eyeballs any consensus is shallow.* In other words, dissent is important in academia and research platforms must be able to accommodate it.

In this paper it was proposed to implement versioning in collaboratories as extended version history. By increasing the significance of version control, research transparency and critical discussion could be improved. The current model of versioning has to be extended in this direction, and the ontological models presented in this article indicate potential ways that this can happen by connecting versions through specific relationships. All of these models draw our focus to the relationships between entities, and by doing so they force us to pay more attention to the events occurring to our information resources and their contributors.

Bibliography

ABO. Annotated books online, edited by Mathijs Baaijens, et al. www.annotatedbooksonline.com. Accessed 31 March 2017.

Albertsen, Ketil, and Van Nuys, Carol. "Paradigma FRBR and Digital Documents." *Functional*

Requirements for Bibliographic Records (FRBR). Hype or Cure-All? edited by Patrick Le Boeuf, Haworth Information Press, 2005, pp. 125–50.

Allman, Eric. "An Introduction to the Source Code Control System", sccs.sourceforge.net/-man/sccs.me.html. Accessed 13 Jan 2019.

Arndt Natanael et al. "Exploring the Evolution and Provenance of Git Versioned RDF Data." *Joint proceedings of the 3rd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2017) and the 4th Workshop on Linked Data Quality (LDQ 2017) co-located with 14th European Semantic Web Conference (ESWC 2017), CEUR Workshop Proceedings*, vol. 1824, edited by Jeremy Debattista, Jürgen Umbrich, and Javier D. Fernández, 2017. ceur-ws.org/Vol-1824/mepdaw_paper_2.pdf. Accessed 3 June 2019.

Baerisch, Stefan. "Versionskontrollsysteme in der Softwareentwicklung." *IZ Arbeitsberichte*, edited by Informationszentrum Sozialwissenschaften der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI), vol. 36, 2005, www.gesis.org/fileadmin/upload/-forschung/publikationen/gesis_reihen/iz_arbeitsberichte/ab_36.pdf. Accessed 26 Aug 2019.

Becker, Hans-Georg. "FRBR, Serials and CIDOC CRM – Modellierung von fortlaufenden Sammelwerken unter Verwendung von FRBRoo."*(Open) Linked Data in Bibliotheken*, edited by Patrick Danowski and Adrian Pohl, De Gruyter, 2013, pp. 64–96.

Blischak, J.D. et al. "A Quick Introduction to Version Control with Git and GitHub." *PLoS Comput Biol*, vol. 12, no. 1, 2016: e1004668. DOI: 10.1371/journal.pcbi.1004668.

Buzzetti, Dino. "Digital Representation and the Text Model." *New Literary History*, vol. 33, 2002, pp. 61–88.

Carusi, Annamaria and Reimer, Torsten. *Virtual Research Environment Collaborative Landscape Study.* JISC, 2010.

Chacon, Scott, and Straub, Ben. *Pro Git.* Git: git-scm.com/book/en/v2/Getting-Started-About-Version-Control. Accessed 3 March 2017.

CIDOC. *Definition of the CIDOC Conceptual Reference Model*, version 6.2 (May 2015). Edited by Patrick Le Boeuf et al. www.cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.pdf. Accessed 3 March 2017.

Danowski, Patrick and Pohl, Adrian, editors. *(Open) Linked Data in Bibliotheken.* De Gruyter, 2013.

Deegan, et al. *Collaborative Research in the Digital Humanities: a Volume in Honor of Harold Short, on the Occasion of His 65th Birthday and His Retirement, September 2010.* Ashgate, 2012.

De Nies, Tom, et al. "Git2PROV: Exposing Version Control SystemContent as W3C PROV." *Proceedings of the ISWC 2013 Posters & Demonstrations Track a track within the 12th International Semantic Web Conference (ISWC 2013), CEUR Workshop Proceedings*, vol. 1035, edited by Eva Blomqvist, and Tudor Groza, 2013.ceur-ws.org/Vol-1035/iswc2013_demo_32.pdf.

IFLA. *Definition of FRBRoo. A Conceptual Model for Bibliographic Information in Object-Oriented Formalism.* edited by Bekiari, Chryssoula et al. IFLA, 2016. www.ifla.org/files/assets/cataloguing/FRBRoo/frbroo_v_2.4.pdf. Accessed 3 March 2017.

Flanders, Julia. "Collaboration and Dissent: Challenges of Collaborative Standards for Digital Humanities." *Collaborative Research in the Digital Humanities*, edited by Marilyn Deegan

and Willard McCarty, Ashgate, 2012, pp. 67–80.

ICARUS. *Monasterium.net.* www.monasterium.net. Accessed 31 March 2017.

Kallass, Kerstin. *Schreiben in der Wikipedia. Prozesse und Produkte gemeinschaftlicher Textgenese.* Springer, 2013.

Lagoze, Carl, and Van de Sompel, Herbert. "The Open Archives Initiative: Building a low-barrier interoperability framework". OAI, 2001, www.openarchives.org/documents/jcdl2001-oai.pdf. Accessed 31 March 2017.

———. *"Compound Information Objects: The OAI-ORE Perspective"* OAI, 2007, www.openarchives.org/ore/documents/CompoundObjects-200705.html. Accessed 24 March 2017.

Le Boeuf, Patrick "A Strange Model Named FRBRoo." *Cataloging & Classification Quarterly*, vol. 50, no. 5–7, 2012, DOI: 10.1080/01639374.2012.679222, pp. 422–38.

———. "FRBR: Hype or Cure-All? Introduction." *Functional Requirements for Bibliographic Records (FRBR). Hype or Cure-All?* edited by Patrick Le Boeuf, Haworth Information Press, 2005, pp. 1–13.

Le Boeuf, Patrick, and David Miller. "Such Stuff as Dreams Are Made On: How Does FRBR Fit Performing Arts?" *Functional Requirements for Bibliographic Records (FRBR). Hype or Cure-All*? Edited by Patrick Le Boeuf, Haworth Information Press, 2005, pp. 151–178.

Meister, Jan-Christoph. "Crowd Sourcing 'True Meaning': A Collaborative Markup Approach to Textual Interpretatio.n" *Collaborative Research in the Digital Humanities*, edited by Marilyn Deegan, Willard McCarty, Ashgate, 2012, pp. 105–122.

Neuroth, Heike, et al. *TextGrid: Von Der Community – Für Die Community: Eine Virtuelle Forschungsumgebung Für Die Geisteswissenschaften.* Verlag Werner Hülsbusch [Printversion] in Kooperation Mit Dem Universitätsverlag Göttingen [Onlineversion] 2015. DOI: doi.org/10.3249/webdoc-3947.

Nyman, Linus. "Understanding Code Forking in Open Source Software. An Examination of code forking, its effect on open source software, and how it is viewed and practiced by developers." Helsinki 2015, helda.helsinki.fi/bitstream/handle/10138/153135/287_978-952-232-275-3-1_v2.pdf?sequence=5&isAllowed=y.

Open Archives Initiative. *ORE Specification – Abstract Data Model,* edited by Carl Lagoze, Herbert Van de Sompel, 2008. www.openarchives.org/ore/1.0/datamodel. Accessed 22 March 2017.

Papyri.info, edited by the Duke Collaboratory for Classics Computing & the Institute for the Study of the Ancient World, papyri.info. Accessed 3 March 2017.

Poe, Marshall. "The Hive." *The Atlantic*, September 2006, www.theatlantic.com/magazine/archive/2006/09/the-hive/305118. Accessed 27 March 2017.

Preston-Werner Tom. "Semantic Versioning 2.0.0." semver.org/. Accessed 13 Jan 2019.

Raymond, Eric. *The Cathedral & the Bazaar.* Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly, 2001.

Robles, Gregorio and Gonzalez-Barahona, Jesus M. "A Comprehensive Study of Software Forks: Dates, Reasons and Outcomes." *Open Source Systems: Long-Term Sustainability*, edited by I. Hammouda et al., OSS 2012. IFIP Advances in Information and Communication Technology, vol 37, Springer, 2012.

Suda On Line: The Byzantine Lexicography. Edited by Stoa Consortium, www.stoa.org/sol.

Accessed 31 March 2017.

Vasold, Gunter. "Progressive Editionen als multidimensionale Informationsräume." *Digital Diplomatics: the Computer as a Tool for the Diplomatist?* edited by Ambrosio, Antonella, et al., Böhlau, 2014, pp. 75–88.

Visser, A.S.Q, and Calis, R.A. "Building a Digital Bookwheel Together: Annotated Books Online and the History of Early Modern Reading Practices." *Bibliothecae.it. Rivista di studi semestrale*, vol. 3, no. 1, 2014, pp. 63–80.

Vogeler, Georg. "Das Verhältnis von Archiven und Diplomatik im Netz. Von der archivischen zur kollaborativen Erschließung." *Digitale Urkundenpräsentationen*, edited by Joachim Kemper and Georg Vogeler, Books on Demand, 2011, pp. 61–82.

Wikipedia contributors. "About." *Wikipedia, The Free Encyclopedia*, 2017, en.wikipedia.org/w/index.php?title=Wikipedia:About&oldid=771132497. Accessed 19 March 2017.

———. "Five pillars." *Wikipedia, The Free Encyclopedia*, 2017, en.wikipedia.org/w/index.php?title=Wikipedia:Five_pillars&oldid=769798419. Accessed 11 March 2017.

———. "Minor edit." *Wikipedia, The Free Encyclopedia*, 2017, en.wikipedia.org/w/index.php?title=Help:Minor_edit&oldid=764768221. Accessed 26 March 2017.

———. "Reliability." *Wikipedia, The Free Encyclopedia*, 2017, en.wikipedia.org/w/index.php?title=Reliability_of_Wikipedia&oldid=772272370. Accessed 26 March 2017.

W3C. *PROV-DM: The PROV Data Model.* edited by Luc Moreau, Paolo Missier, W3C, 2013. www.w3.org/TR/2013/REC-prov-dm-20130430. Accessed 18 March 2017.

Appendix

Biographical Notes

Gioele Barabucci is a Marie Curie Experienced Researcher at the Cologne Center for eHumanities of the University of Cologne. He received his PhD in Computer Science from the University of Bologna. His main research topics are the design of knowledge (how to represent and store information) and the evolution of information (understanding and forecasting how data and its structure will change over time). Concretely, this means studying comparison algorithms, devising versioning systems, formalizing document models, as well as researching ontologies, legal documents and multilingual systems.

He is currently working on the formalization of the concept of 'document evolution': how documents change through time due to human edits, changes of format, translations, and so on. In the past he has worked on collation systems for large-scale critical editions (Capitularia, Averroes), the Cologne Sanskrit Dictionary, the Akoma Ntoso standard for legal documents and many other academic and open source projects.

Roman Bleier studied History and Religious Studies at the University of Graz. Following his studies in Graz, he worked on the *Saint Patrick's* Confessio *HyperText Stack*, a project of the Royal Irish Academy in Dublin, and completed a Ph.D. in Digital Arts and Humanities at Trinity College Dublin. After finishing his Ph.D. in 2015, he was CENDARI Visiting Research Fellow at King's College London and worked at An Foras Feasa, the Digital Humanities Centre at Maynooth University. In 2016 he returned to Graz as DiXiT Fellow and has since worked as postdoctoral researcher on different digital scholarly editing projects at the Centre for Information Modelling – Austrian Centre for Digital Humanities.

Richard Breen holds a Master's degree in Digital Humanities from Maynooth University, as well as an International Bachelor of Music degree from Maynooth University. He has previously worked on the *Letters of 1916* as well as developing the *Concorde40* archive at the Contemporary Music Centre in Fishamble Street, Dublin. He developed Documenting Transmission: The Rake Cycle as part of his Master's Thesis at Maynooth University.

Richard currently works in the Centre for Environmental Humanities at Trinity College Dublin as a Research Assistant on the *The North Atlantic Fish Revolution: An Environmental History of the North Atlantic 1400-1700 (NorFish)* project. His current research interests are Digital Humanities, Musicology, and History.

Martina Bürgermeister holds a Master's degree in History and a Master's degree in Digital Humanities. Since 2013, she has been working for the Centre for

Information Modelling – Austrian Centre for Digital Humanities at the University of Graz. In 2014, she started co-developing the platform monasterium.net within the project *Illuminierte Urkunden als Gesamtkunstwerk*.

Elisa Nury is a postdoctoral researcher at the University of Geneva for the Grammateus project on Greek documentary papyri. In 2018, she completed a Ph.D. in Digital Humanities at the University of King's College London, UK, on the topic of automated collation tools and digital critical editions. She graduated from the University of Lausanne, Switzerland, with a specialisation in History of the Book and Critical Edition. Her research interests include Latin literature, digital humanities and digital scholarly editing.

Nicholas Pickwoad trained in bookbinding and book conservation with Roger Powell, and ran his own workshop from 1977—89. He has been an Advisor on book conservation to the National Trust since 1978. He was Chief Conservator in the Harvard University Library from 1992–95 and is now project leader of the St Catherine's Monastery Library Project based at the University of the Arts London, where he is director of the Ligatus Research Centre, which is dedicated to the history of bookbinding. He lectures and teaches extensively on the history of European bookbinding in Europe and the USA.

Martina Scholger is a senior researcher at the Centre for Information Modelling – Austrian Centre for Digital Humanities at the University of Graz. She studied art history and received her PhD in digital humanities with a study on artists' creation processes documented in their notebooks. She is teaching information modelling and is involved in digital scholarly editing projects. She has been a member of the Institute for Documentology and Scholarly Editing since 2014 and a member of the TEI Technical Council, where she is currently serving as Chair, since 2016.

Christian Thomas has studied German Literature and Philosophy at the Humboldt-Universität zu Berlin (HU). Since 2010 he is working as a research assistant at the Berlin-Brandenburg Academy of Sciences and Humanities for the projects *Deutsches Textarchiv* (DTA) and *CLARIN-D*. Additionally, from 2014–16 he has been working as a research assistant in the *Hidden Kosmos* project, funded by the Excellence Initiative at the HU. He is currently working on his doctoral thesis on the topic of A. v. Humboldt's Kosmos-Lectures.

Athanasios Velios is Reader in Documentation at Ligatus, University of the Arts London. He studied archaeological conservation at the Technological Educational Institute of Athens and he completed his PhD at the Royal College of Arts, London on computer applications to conservation. He works with Linked Data

technologies to model records from memory institutions. He is a member of the CIDOC-CRM special interest group and he served as webmaster of the IIC (2009-2018). He works closely with Professor Nicholas Pickwoad on the documentation of binding structures, including the development of an XML schema for binding records and the publication of the Language of Bindings Thesaurus.

Georg Vogeler is Professor for Digital Humanities at the University of Graz. He studied Historical Auxiliary Sciences in Freiburg (Brsg.) and Munich. In 2002 he received his PhD with a study on late medieval tax registers of German territories and in 2016 the venia docendi for his Habilitationschrift on the use of the charters of Frederic II (1194-1250) by his contemporaries in Italy.

His research interests are in the field of late medieval administrative records, the diplomatics of the charters of Frederic second, digital diplomatics, digital edition and the application of semantic web technologies to humanities research questions. He has received several international grants in these fields.

Sean M. Winslow is a postdoctoral researcher at the Centre for Information Modelling – Austrian Centre for Digital Humanities of the University of Graz. He is a is a historian of manuscript practices, focusing on medieval Europe and Ethiopia. His current work includes the digital modelling of medieval charters, the description of Ethiopian binding decoration, and the digital cataloguing of Syriac manuscripts.