Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

Schriften des
Instituts für Dokumentologie und Editorik

herausgegeben von:

Band 15

Schriften des Instituts für Dokumentologie und Editorik — Band 15

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing

edited by

Elena Spadini, Francesca Tomasi, Georg Vogeler

2021

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 5. Dezember 2021.

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7543-4369-2
Einbandgestaltung: Stefan Dumont nach Vorarbeiten von Johanna Puhl und Katharina Weber
Satz: LuaTEX, Bernhard Assmann

Contents

Projects and Editions

Appendices

Infrastructures and Technologies

Formal Models

Projects and Editions

Modelling Cross-Document Interdependencies in Medieval Charters of the St. Katharinenspital in Regensburg

Colin Sippl, Manuel Burghardt, Christian Wolff

Abstract

To overcome the limitations of structural XML mark-up, *graph-based data models* and graph databases, as well as *event-based ontologies* like *CIDOC-CRM* (FORTH-ICS 2018) have been considered for the creation of *digital editions*. We apply the graph-based approach to model charter regests and extend it with the CIDOC-CRM ontology, as it allows us to integrate information from different sources into a flexible data model. By implementing the ontology within the *Neo4j* graph database (Neo4j 2018) we create a sustainable data source that allows *explorative search queries* and finally, the integration of the database in various technical systems. Our use case are the charters from the *St. Katharinenspital*, a former medieval hospital in Regensburg, Germany. By analysing charter abstracts with *natural language processing (NLP)* methods and using additional data sources related to the charters, we generate additional metadata. The extracted information allows the *modelling of cross-document interdependencies of charter regests* and their related entities. Building upon this, we develop an *exploratory web application* that allows to investigate a graph-based digital edition. Thereby, each entity is displayed in its unique context, i.e., it is shown together with its related entities (next neighbours) in the graph. We use this to enhance the result lists of a full-text search, and to generate entity-specific detail pages.

1 Introduction

The creation of digital editions is one of the main applications in the Digital Humanities. In recent years, the use of the eXtensible Markup Language (XML) along with Text Encoding Initiative (TEI) styles has become the most common way to do this (Sahle 2013, 341–42). However, Kuczera (2016) and Sahle (2013, 345–71) point out some severe limitations of digital editions that merely rely on structural XML mark-up. Among the main drawbacks of XML is its inability to model overlapping structures or parallel annotation hierarchies (Kuczera 2016). To overcome these limitations, graph-based data models and graph databases (Kuczera 2017), as well as event-based ontologies like *CIDOC-CRM* (Le Bœuf et al. 2015), have been considered for the creation of digital

Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing. Ed. by Elena Spadini, Francesca Tomasi and Georg Vogeler. Schriften des Instituts für Dokumentologie und Editorik 15. Norderstedt: Books on Demand, 2021. 181–203.

editions (Ore 2009). The graph-based approach has already been tested with charters of the *Monasterium.Net*[1] platform (Jeller 2019). However, existing implementations are highly experimental and only highlight basic aspects of the graph-based approach, e.g., the ability to analyse mutual relationships as well as new visualisation types. Finally, these use cases are unsuitable for end users, as they are just proof-of-concepts. We apply the graph-based approach of modelling digital editions and extend it with the CIDOC-CRM ontology, as it allows us to integrate information from different sources into a flexible data model. We achieve this by preparing raw data in such a way that charters from our data source can easily be linked together in a graph database. By using the CIDOC-CRM ontology, we create a sustainable data source that allows for specific and explorative search queries and, finally, integration into various technical systems. For this purpose, we rely on the *Neo4j graph database*.[2] Thereby we are shifting the focus from the structure of the documents to their *entities* and *relations*. Our use case are charters from the *St. Katharinenspital*, a former medieval hospital in Regensburg, Germany. These charters bear witness to changes in the power structure in Regensburg over the course of several centuries (Kaufner 2011, 13–21). In the following, we give an overview of our dataset, the process of data preparation, data analysis, as well as our data model. Building upon this, we develop an exploratory web application[3] that can be used to investigate a graph-based digital edition of charters from the St. Katharinenspital. Our implementation serves as a blueprint for the future development of a digital repository for graph-based scholarly editions of medieval charters.

2 The St. Katharinenspital Dataset

In total, the archive of the St. Katharinenspital holds more than 4,000 charters, which is a considerable number for a rather small archive (König 2003, 16). Hence, the archive contains an essential part of the historical written heritage of Regensburg and its surrounding area. The dataset from which the entities and relations are extracted consists of three main components: the *CEI-XML*[4] charter regests on Monasterium.Net, the scholarly editions series *Die älteren Urkunden des St. Katharinenspitals in Regensburg* and other works,[5] as well as a dataset of different files from the archives of

---

1 St. Katharinenspital dataset on Monasterium.Net, available at http://monasterium.net/mom/DE-AKR/Urkunden/fond, all hyperlinks in this article were last accessed on Oct. 7^{th}, 2019.

2 Neo4j Graph Database (Neo4j 2018).

3 Currently only available in German language, available at https://urkunden.ur.de.

4 Charters Encoding Initiative, CEI – The Project. Mark-up for medieval and early modern legal records (Vogeler 2004).

5 *Die älteren Urkunden des St. Katharinenspitals in Regensburg* currently consists of three printed volumes (König 2003; Kaufner 2011; Sturm 2013). Additionally, we incorporate an unpublished volume (Feichtmeier, n.d.) and an earlier work about medieval chancery in Regensburg (Ambronn 1968).

St. Katharinenspital, including MS Office files, scans and digital documents containing the historical tradition of the St. Katharinenspital charters. Currently, 1,050 charter regests are already available on Monasterium.Net as CEI-XML documents. These documents are of varying lengths, levels of detail, and represent different states of primary source analysis. The documents contain 712 places (292 distinct) and 1,217 (749 distinct) persons as manually annotated entities.[6] Due to missing links between the CEI-XML documents and a lack of standardisation, uniform spellings of those entities are largely missing. Furthermore, a mixture of entity names with additional (e.g., biographical) data frequently[7] occurs:

```
<cei:back>
  <cei:persName>Karl der Große, fränkischer König und Kaiser</cei:persName>
  <cei:placeName>Frankfurt a. Main (krfr.St., Hessen)</cei:placeName>
  <cei:placeName>Vivarias (Gewässer bei Regensburg)</cei:placeName>
  <cei:placeName>Pielmühle (Gde. Lappersdorf, Lkr. Regensburg)</cei:placeName>
</cei:back>[8]
```

The example shown above highlights a structural problem of documents annotated with flexible data schemes such as CEI-XML. Since common goals for advanced annotation and analysis of primary documents can be achieved in different ways, ambiguities may occur among the individual documents. The exact spelling of place names or person names (e.g., "Karl der Große"), along with any normalisation efforts, are ultimately left to the editor of a CEI-XML file on Monasterium.Net (Jeller 2019). Because of these varying spellings and structural differences between single CEI-XML files in the dataset[9], linking and analysing the documents is severely impeded. Additionally, for the St. Katharinenspital dataset there are no finding aides, like indices or registers, online. Moreover, archive IDs are inconsistently distributed among our Monasterium.Net dataset, and personal identifier (PID) references to authority files like the *Gemeinsame Normdatei* (Integrated Authority File, GND: Deutsche Nationalbibliothek 2018) are missing. Together with the limited full text search capabilities of Monasterium.Net, the issues related to the St. Katharinenspital dataset pose restrictions to working with the CEI-XML data as well as linking them to external data sources. Besides, an additional 1,699 regests (as of 2018) have been transcribed or extended by means of Microsoft Office tools (Word, Excel and Access) and are part of the data collection of the St. Katharinenspital. The complete transcription of *Repertorium C*, known as *DicMihi*, is one of the most important file sources. The original source is a register from 1745, which lists the St. Katharinenspital charters

---

6 These numbers were determined by XQuery queries and after an extensive data cleansing process.

7 E.g., "Otto Prager (1243–1244, 1248–1251, 1255)" or "Eberhard, Graf von Abensberg, Erzdiakon".

8 München, Bayerisches Hauptstaatsarchiv Kloster St. Emmeram Regensburg Urkunden (0794-1800) BayHStA, Kloster St. Emmeram Regensburg Urkunden 1, available at http://monasterium.net/mom/DE-BayHStA/KURegensburgStEmmeram/000001/charter.

9 I.e., inconsistent usage of XML tags.

Salbuch Nordgau		
charter no.	**archive ID**	**Monasterium.Net ID**
3	SpAR Urk. 1479	12510404
4	SpAR Urk. 54	12510907
9	SpAR Urk. 1074	12530626
14	SpAR Urk. 1073	12540000
21	SpAR Urk. 395	12550803
25	SpAR Urk. 1288	12560515
26	SpAR Urk. 55a	12560617

Table 1. Sample of the tradition of the charters in urbarium *Salbuch Nordgau* listed by Kaufner (2011). Charter numbers originate from the scholarly edition.

according to places (König 2003, 23). It provides a whole set of categories[10] and places. These categories and places can be used to assign attributes to those charters which are listed in the finding aide. Additionally, the transcription of *DicMihi* lists charters that are now lost. Apart from the *DicMihi* register, the tradition of the St. Katharinenspital charters listed in various scholarly editions was also added to the data source. Initially, all documents from the printed editions were added to a temporary MySQL database with their charter number, archive IDs and Monasterium.Net IDs. Table 1 shows which of the charters in Kaufner (2011) are contained in the *Salbuch Nordgau* urbarium with their corresponding IDs.

In total, traditional references from five different scholarly editions covering 176 St. Katharinenspital charters were added to the dataset. Hence, our dataset now contains information about the tradition of charters in various historic and scholarly documents. We use this information to model the traditional context of the documents in the graph database. In conclusion, our dataset of the St. Katharinenspital is particularly suitable for the extraction of entities and relations to establish links between the charters and related documents. By using this dataset, we are already able to add a significant amount of knowledge to the rather outdated St. Katharinenspital charters collection on Monasterium.Net.

3 Extracting Entities and Relations

The heterogeneous St. Katharinenspital dataset provides the entities and relations to model cross-document interdependencies. We extract these by identifying mutual

[10] E.g., "fürstliche Privilegia", "die Spitalmühl und das Baad betreffende Brief" or "Bischöffliche Begnadigung und andere Urkunde".

entities in different individual documents, as well as by adding additional information, e.g., norm data,[11] via data normalisation efforts and natural language processing (NLP) with *spaCy*[12]. Finally, the enhanced data are modelled as an interconnected graph-structure that can be used for exploratory analyses of both the documents and the entities from many different research perspectives (e.g., historical, archival, linguistic, economic and cultural). The following types of entities are extracted or determined during the data pre-processing and analysis:

- charters (individual regests, transcripts, references…);
- legal activities (legal content of a charter…);
- actors (authors, witnesses, groups, legal entities…);
- related documents (scholarly editions, traditional documents…);
- dates (time stamps); and
- places (with or without geo-reference).

To build the data collection, the raw data are processed in an extensive data pre-processing pipeline, which consists of several individual operations. Primarily, the data are cleaned, labelled and restructured to make them machine-readable by generating *.csv, *.html or *.xlsx files. Thereby, the identified entities and relations get updated manually and automatically to be able to uniformly acquire them later on setting up the graph database. This part of the whole process is particularly time-consuming, as almost all raw data need to be sighted. Following this, normalised spellings of the entities and PIDs (archive IDs, GND references, file names etc.) are introduced. Subsequently, parts of the collected entities and relations are stored in a temporary MySQL database. Thus, the data are quickly available at different points of the preparation process, e.g., for additional metadata aggregation, such as geoparsing,[13] further data cleaning processes, and, finally, to set up the graph database (see Figure 1). Since the extraction of the entities and relations from the St. Katharinenspital dataset is experimental and the final quality of the results could not be anticipated, we decided not to store the data in an RDF-like data structure, and kept on using the MySQL database during the whole data preparation process (more on this in the following section).

We see charter abstracts as a valuable resource for the extraction of entities and relations. Therefore, we particularly focus on their analysis. Thereby, the NLP methods applied to extract entities and relations from the charter abstracts form an integral part of our data pre-processing pipeline. Charter abstracts are just a brief summary of

[11] We incorporate experimentally the Integrated Authority File (GND), managed by the German National Library (DNB), available at http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html.

[12] We use spaCy (Honnibal & Montani 2018) to extract linguistic features like part-of-speech tags, dependency labels and named entities from our corpus. Available at https://spacy.io/usage/linguistic-features.

[13] In this project, basic geo data are retrieved from GeoNames database (GeoNames 2018), in a process comparable to Jeller (2019).

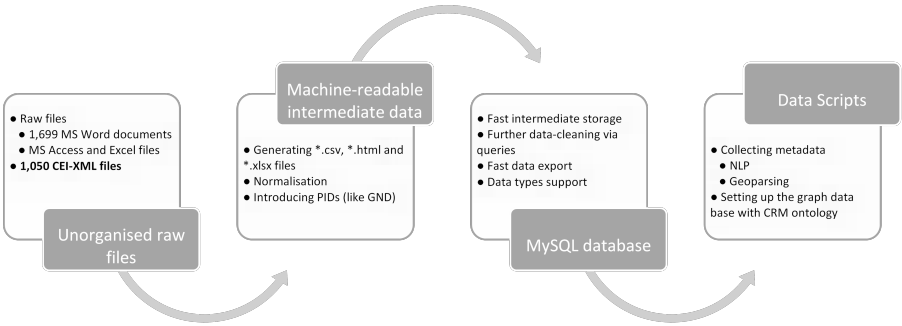

Figure 1. Data pre-processing pipeline. Illustration created by the authors.

the primary content of a charter and often consist of a single sentence. The following sentence is a typical example:

> "Ulrich der Frutrunch von Abbach verkauft dem St.-Katharinenspital sein Gut in Teingen (Teugn) um 15 Pfund Regensburger Pfennig."(St. Katharinenspital charter SpAR Urk. 2101)[14] *("Ulrich der Frutrunch von Abbach sells his manor in Teingen (Teugn) to the St.-Katharinenspital for 15 pounds of Regensburger Pfennig.")*

The charter abstracts are rather uniform, and thus facilitate information extraction. The approach of collecting data from German charter abstracts was already applied by Kuczera (2017): by extracting the first verb of a charter abstract and applying a lemmatizer afterwards, Kuczera creates basic lemma nodes in the graph database. The nodes are connected to the regest nodes via *HERRSCHERHANDELN* ("acts-of-rulership") edges (2017, 187). The particular focus on the verb essentially follows the theory of *dependency grammar*, according to which the dependence of a word upon another word results from a verb. This grammatical concept goes back to Lucien Tesnière (2015), and views the finite verb as the centre of the grammatical organisation of a sentence. There also exists a basic NLP script for spaCy by Georg Vogeler 2018),[15] which can be used to analyse a corpus of regests. The script automatically assigns attributes to digital regests by language determination and basic named entity extraction (NER). This script shows how relatively simple NLP techniques

[14] Regensburg, Archiv des Katharinenspitals Urkunden (1145-1568) SpAR Urk. 2101, available at https://www.monasterium.net/mom/DE-AKR/Urkunden/SpAR_Urk_2101_/charter.

[15] This is a collection of scripts and workflows for NLP experiments with data from Monasterium.Net, available at https://github.com/GVogeler/mom-NLP.

can be used to evaluate large amounts of charter regests of Monasterium.Net. More complex syntactic structures, however, are not evaluated by the script. We extend both Kuczera's and Vogeler's NLP-based approaches with a heuristic analysis of sentence structures found in German charter abstracts of the St. Katharinenspital. Essentially, we apply the spaCy default NLP pipeline (a tokenizer, a part-of-speech tagger, a dependency-parser and named entity recognition[16]) with some small modifications, such as merging identified multi-token named entities (NEs) into single tokens.[17] Analysing a charter abstract (e.g., of charter SpAR Urk. 2101[18]) with spaCy generates a *spaCy Doc object.*[19] It includes a sequence of part-of-speech tagged tokens and syntax dependencies that can be represented as follows (Figure 2):

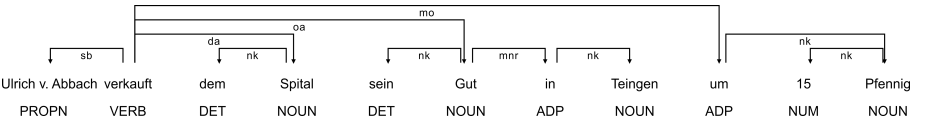

Figure 2. SpAR Urk. 2101 charter abstract syntax dependencies. Illustration created by the authors with *displaCy*[20].

In the following, a heuristic is derived from the generated dependency tree to extract the entities contained in the German language regests and their relationships to each other. The heuristic approach focuses on extracting *triples* (SVO) or *quadruples* (SVOO) of entities and verbs (nodes) as well as their mutual relations (edges) from the charter abstracts. The relations are subsequently derived from the word order of the single elements of a triple or quadruple:

```
S = subject, V = verb, O = object

SVO:['Gebolf von Ellenbach', 'verkaufen (to sale)', 'Weingut (vineyard)']
S-[:relationA]-V-[:relationB]-O

SVOO:['Rudiger der Mulnar', 'vermachen (to devise)', 'St.Katharinenspital', '
     Äcker (fields)']
S-[:relationA]-V-[:relationB]-O-[:relationC]-O
```

[16] SpaCy language processing pipelines, available at https://spacy.io/usage/processing-pipelines.

[17] We collapse consecutive tokens tagged as a named entity into a single token. Thus, a multi-token entity like "Ulrich von Abbach" that consists of tokens belonging to different lexical categories now becomes a single token with a single part-of-speech tag.

[18] SpAR Urk. 2101 on Monasterium.Net, available at https://www.monasterium.net/mom/DE-AKR/Urkunden/SpAR_Urk_2101_/charter.

[19] SpaCy Doc object definition, available at https://spacy.io/api/doc.

[20] DisplaCy Dependency Visualizer is spaCy's visualisation tool. An online demo is available at https://explosion.ai/demos/displacy.

Based on the extracted dependencies, further information may be added to the direct and indirect objects, e.g., if a direct object can be identified as a *E53Place* CRM class.[21] The heuristic is experimental and inevitably will create false positives. It heavily depends on the language model used by spaCy, as well.[22] Our heuristic basically performs four tasks for all tokens in the charter abstract. A simplified representation of the heuristic code can be seen below and is also available on GitHub:[23]

```
# heuristic to extract quadruples and triples from German charter abstracts
index = 0
subject = ''

dobject = ''
dobject2 = ''
# iterates through every token of spaCy doc object
for token in doc:
  if token.dep_ == "sb" or token.dep_ == "oa" or token.dep_ == "da":
    # task 1: get subject
    if token.dep_ == "sb" and index == 0:
      subject = token.text
    # task 2: get verb of subject (predicate)
    if token.dep_ == "sb" and index == 0 and token.head.pos_== "VERB":
      verb = token.head.lemma_
    # task 3: get direct object
    if index == 1 and (token.dep_ == "da" or token.dep_ == "oa")
                  and token.head.pos_ == "VERB":
      dobject = token.text
    # task 4: get indirect object
    if index == 2 and token.dep_ == "oa":
      dobject2 = token.text
    index += 1
if subject != '' and verb != '' and dobject != '' and dobject2 != '':
  return subject, doc, verb, dobject, dobject2
if subject != '' and verb != '' and dobject:
  return subject, doc, verb, dobject
return ''
```

The sample code shows the structural criteria that allow us to identify a structural subject, verb and object(s) in a simple German declarative phrase (Meibauer et al. 2013, 20–50). Therefore, the success of the extraction depends on the syntax of the charter abstract.[24] For the German language, we start from the *V2 word order*, which places the finite verb of a phrase or sentence in second position, with a single constituent preceding it (Haider 2010, 1f.). By using the syntax dependencies, the part-of-speech

[21] We achieve this by matching words and performing further syntactic analyses that would go beyond the scope of this publication.

[22] SpaCy's pre-trained statistical models for German, available at https://spacy.io/models/de.

[23] The implementation used is available on GitHub under GPL 3 license and can be used to generate an experimental Neo4j database with some German charter abstracts (Sippl 2019), available at https://github.com/cs-ubr/charter-abstracts.

[24] Charter abstracts in our dataset vary in length and detail. A significant part of the corpus consists of abstracts that consist of more complex syntactic structures with several phrases or even several sentences.

Dependency Label[25]	Description	POS	Token	CRM Class
sb	subject	PROPN	Ulrich v. Abbach	E21Person
-	(predicate)	VERB	verkauft (*sells*)	E7Activity
da	dative	NOUN	Spital (*St. Katharinenspital*)	E74Group
oa	accusative object	NOUN	Gut (*manor*)	E53Place

Table 2. Quadruple of subject, predicate, direct and indirect object for the charter abstract of SpAR. Urk. 2101.

tags and the word order from the spaCy Doc object, we are now able to derive the constituents of a triple or a quadruple of a phrase. For the subject and the verb this is implemented in lines 10 to 15 in the sample code. Just like Kuzcera (2017), we also use the lemmatised form of the verb. However, we use the verb to generate an entity instead of a relation (see next section). The extraction of direct and indirect objects is implemented in lines 16 to 21. If no indirect object could be extracted, only a triple consisting of the subject, the verb and the direct object is returned. Also, different spellings of actors or place names may appear in the charter abstracts. For example, *St. Katharinenspital* may occur as *St.-Katharinenspital*, *Spital*, or simply *Katharinenspital*. We therefore introduce normalised spellings for selected entities (e.g., *St. Katharinenspital*). Table 2 shows the extracted entities of the example charter that are already assigned with their corresponding CRM classes.

Table 3 shows a few example quadruples extracted from charter abstracts. The entities shown can be linked to a charter and the identified dependencies also help to describe the relationships between the entities. As can be seen, the lemmatised verbs particularly stand out, as they make it possible to categorise the charters by their legal content. The results also show that triples and quadruples are suitable to generate human readable lists. Thus, improving the data can be achieved quickly, either manually or automatically. Our heuristic identifies 225 quadruples and 407 triples in the charter data, thereby raising the number of annotated places from 292 to 1351, as well as the number of persons from 749 to 2435[26] (see Table 4 and Table 5). The results could be improved if the machine learning features of spaCy, and more detailed heuristics for different syntactic structures, were applied. The time and effort required for this, however, is too big for the collection of a small archive. In the case of much larger corpora of charter abstracts (e.g., the whole Monasterium.Net corpus),

[25] The spaCy dependency labels are based on the TIGER Treebank annotation scheme, available at https://spacy.io/api/annotation#dependency-parsing.

[26] The numbers show distinct entities. However, ambiguities and false positives remain in the data and thus, in the numbers.

Charter	grammatical subject	predicate	direct object	indirect object
SpAR_Urk_799	Rudger der Mulnar	vermachen (*to bequeath*)	St. Katharinenspital	Äcker (*fields*)
SpAR_Urk_135	Leopold von Gründlach	verleihen (*to give*)	St. Katharinenspital	Ablass (*indulgence*)
SpAR_Urk_1394	Romungus von Chamerstein	bestätigen (*to confirm*)	St. Katharinenspital	Besitz (*property*)
SpAR_Urk_1367	Otto von Dürn	verkaufen (*to sell*)	St. Katharinenspital	Hof (*farm*)
SpAR_Urk_1189	Pernger von Haydawe	übereignen (*to transfer*)	St. Katharinenspital	Wald (*forest*)
SpAR_Urk_1072a	Elspet	verkaufen (*to sell*)	St. Katharinenspital	Teil (*part*)

Table 3. Quadruples extracted from the charter abstracts by means of NLP.

Property	Value
Charters (CEI-XML files)	1,050
Persons (unique)	749
Places (unique)	292

Table 4. Entities in St. Katharinenspital Monasterium.Net dataset.

this would be a feasible approach. Additionally, our extracted data is suitable to train a *Conditional Random Fields* (CRF) classifier for improved NER results and even an automatic assignment of CRM labels. This also allows for a context sensitive extraction of numbers (e.g., quantifiers, amounts of money etc.) and thereby substantially increase the amount of extracted information. Therefore, this approach may be considered in future work. An example of a successful use of a pre-trained CRF classifier in a comparable DH context with extensive data preparation can be found in Lüschow (2020).

4 Modelling Charters and Cross-Document Interdependencies

To model the St. Katharinenspital dataset as a graph, we rely on two essential concepts: entities (nodes) and relations (edges). We use the Neo4j labelled-property graph

DB Property	Value
Nodes total	9,676
CRM-Entities (E1CRMEntity)	7,665
Edges total	24,138
Charters (E5Event)* *E7Event excluded	2,489
Persons (E21Person)	2,435
Places (E53Place)	1,351

Table 5. Entities and properties of the final labelled-property graph database.

database to store these entities and relationships. Compared to *RDF triplestores*, Neo4j supports advanced graph metrics, weighted edges and is very easy to set up and maintain from a developer's perspective. The first two aspects are relevant for network analyses, which are becoming increasingly important in DH (Jannidis et al., 2017, 147-149). The latter facilitates, as an example, the future integration of Neo4j into a large digital repository based on software for research data management such as Invenio.[27] This way, users and developers can also get around the major disadvantages of SPARQL (Vogeler 2019). In the academic world, however, RDF triplestores are widely used for storing and retrieving triples with semantic queries. A recent example for this is the project of Lüschow (2020). Therefore, the migration of our data into an RDF database, after an extensive data evaluation process, is a future application. However, a detailed comparison of RDF and Neo4j would go beyond the scope of this paper, especially since it is also the subject of an ongoing debate[28].

The extracted entities can take various forms, including persons, institutions or places. Kuczera's (2017) example of verb extraction and the modelling of relationships between regests and persons as *HERRSCHERHANDELN* ("acts-of-rulership") edges in the graph is ultimately a description of a historic event. According to our interpretation, the extracted verb describes this abstract event, which in turn can be assigned to a date, a person, an object that is changed by the event and a specific document (Figure 3).

Events are a central concept in modelling data from domains such as history or cultural heritage (Van Hage et al. 2011). However, the granularity of a data model,

[27] Invenio is an Open Source framework for large-scale digital repositories. It was initially developed by CERN and features support for large-scale research data managmenet use cases, available at https://invenio-software.org/.

[28] E.g., this article on the technology news website ZDNet contrasts contrary statements of Neo4j's CEO with the statements of an advanced GraphDB user, thus highlighting the technological and ideological differences between the two approaches, available at https://www.zdnet.com/article/graph-databases-and-rdf-its-a-family-affair/.

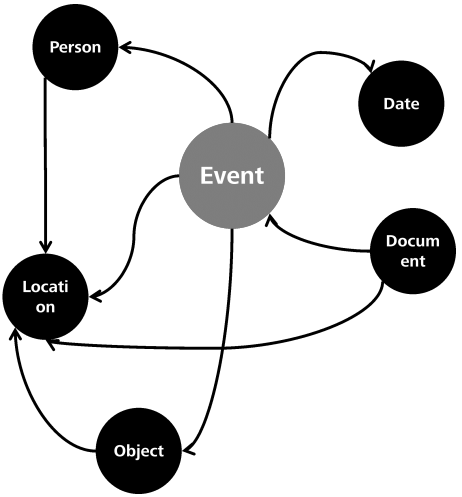

Figure 3. Event as central element of a graph-based data model. Illustration created by the authors.

the data to be captured, and the essential components of it, ultimately depend on the overall requirements a data model needs to meet. As we describe relations between entities by incorporating and linking information from CEI-XML files, Microsoft Office files, archival metadata, biographical data, place attributions, roles, basic prosopographic knowledge or data we generate with NLP, we require a data model that allows us to map basic forms of traditional diplomatic work. This data model needs to be sustainably implemented in a versatile technical environment, as users (institutions, students, scholars and software developers) have different demands. CIDOC-CRM is a model that can be used within graph databases, and it is particularly suitable if complex problems can't be solved in a monodisciplinary manner. It is an ISO standard and intended to enable discipline-independent linking of heterogeneous cultural information, especially from cultural heritage stored in museums, libraries or archives. For this purpose, logically defined terms – classes and properties – are used. Both are intended to support a great number of information resources (Le Bœuf et al. 2015, ii). Apart from using the ontology for medieval charters (Ore 2009), there have been proposals to use CIDOC-CRM for account books as well (Vogeler 2015), thus further primary sources from the St. Katharinenspital archive (e.g., account books or urbaria) can easily be added to the model in future applications. A factoid-based ontology expressed by the CIDOC-CRM model was also proposed to improve interoperability, and enable better support of prosopographic databases (Pasin & Bradley 2015). This outlines the fact that very different source types can be incorporated in a single data

Nodes (CRM classes)	Edges (CRM properties) connected to E5Event
E7Activity	P20HadSpecificPurpose P9ConsistsOf
E21Person	P11HadParticipant
E30Right	P129IsAbout
E31Document	P70Documents
E52TimeSpan	P4HasTimeSpan
E53Place	P7TookPlaceAt P161HasSpatialProjection
E55Type	P2HasType
E74Group	P11HadParticipant

Table 6. Edges of *E5Event* in the data model.

model, and how well it is suited to the project. Central classes of the ontology are *E5Event*[29], thing *E70Thing*[30] and *E39Actor*[31]. To better illustrate how the CRM classes can be extracted from charter abstracts, they are annotated in the following for charter SpAR Urk. 2101:

> [E21Person: *Ulrich der Frutrunch von Abbach*] [E7Activity: *verkauft (sells)*] dem [E74Group: *St.-Katharinenspital*] sein [E53Place: *Gut (manor)*] in [E53Place: *Teingen (Teugn)*] um 15 Pfund Regensburger Pfennig. (St. Katharinenspital charter SpAR Urk. 2101)

As described in the previous section, entities and relations are derived from the syntactic structure of the regests abstracts using our spaCy NLP heuristic. The resulting data model is based on the CIDOC-CRM v6.2.1. Thereby, the following CRM relationships are used to connect a charter (*E5Event*) to other CRM entities (Table 6):

The nodes (*E1CRMEntities*) in our data model may be connected via different edges to a charter (*E5Event*). The easiest way to understand this is by looking at the extracted quadruples. A charter as an instance of CRM class *E5Event* is at the centre of our data mode. As a result, the quadruple for the example charter consists of `"SpAR Urk. 2101"` (*E5Event*), `"Ulrich v. Abbach"` (*E21Person*), `"verkaufen"` *E7Activity*), `"Spital"` (*E39Group*), and `"Gut"` (*E53Place*) class instances, and is represented within the graph database as illustrated in Figure 4.

[29] "This class comprises changes of states in cultural, social or physical systems, regardless of scale, brought about by a series or group of coherent physical, cultural, technological or legal phenomena." (Le Bœuf et al. 2015, 5).

[30] "This general class comprises discrete, identifiable, instances of *E77PersistentItem* that are documented as single units, that either consist of matter or depend on being carried by matter and are characterized by relative stability." (Le Bœuf et al. 2015, 70).

[31] "This class comprises people, either individually or in groups, who have the potential to perform intentional actions of kinds for which someone may be held responsible." (Le Bœuf et al. 2015, 20).

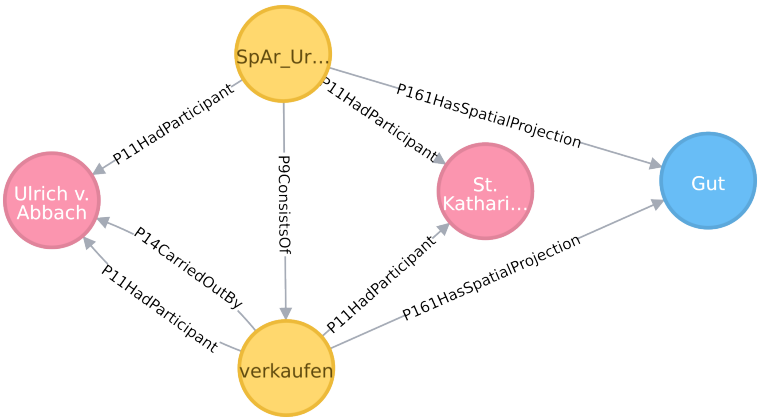

Figure 4. Graph representation of charter abstract SpAR Urk. 2101. Illustration created by the authors with Neo4j Browser.

The visualisation tool of the Neo4j data base allows for the easy creation of graph views, thus the overall properties of the generated data model can be quickly checked. The nodes (entities) are connected via various edges that represent different semantic relationships. Since the example given is a *property sale*, there is a *seller* (Ulrich v. Abbach) and a *buyer* (St. Katharinenspital). They participate in the transaction *E7Activity*, so both seller and buyer are connected via a *P11HadParticipant*[32] CRM property to the *E7Activity*. As the charter is represented as an instance of *E5Event*, both actors are also connected to it via a *P11HadParticipant* edge. As an abstract event, the charter describes the property sale. Hence, it consists of (*P9ConsistsOf*[33]) the property sale. This logic can be applied to all relationships depicted in Figure 4. If all nodes are connected correctly, five nodes with eight edges are created in the database from just a single quadruple. The CIDOC-CRM constraints and structure are enforced with the *cidoc-crm-neo4j* script[34], so that every CRM subclass has the labels of its super classes, when created in the database. Also, the validity of the data model is guaranteed by enforcing the right relations between individual nodes and inhibiting relations that are contrary to the CIDOC-CRM definition. Hence, the

[32] "This property describes the active or passive participation of instances of E39 Actors in an E5 Event." (Le Bœuf et al. 2015, 48).

[33] "This property associates an instance of E4 Period with another instance of E4 Period that is defined by a subset of the phenomena that define the former." (Le Bœuf et al. 2015, 47).

[34] The CIDOC-CRM constraints and class hierarchies are enforced with *cidoc-crm-neo4j*, a python-based script by Erick Peirson, ASU, GPL3 (2014), which in turn relies on the latest RDFS serialisation of CIDOC-CRM (v6.2.1), available at http://www.cidoc-crm.org/sites/default/files/cidoc_crm_v6.2.1-2018April.rdfs.

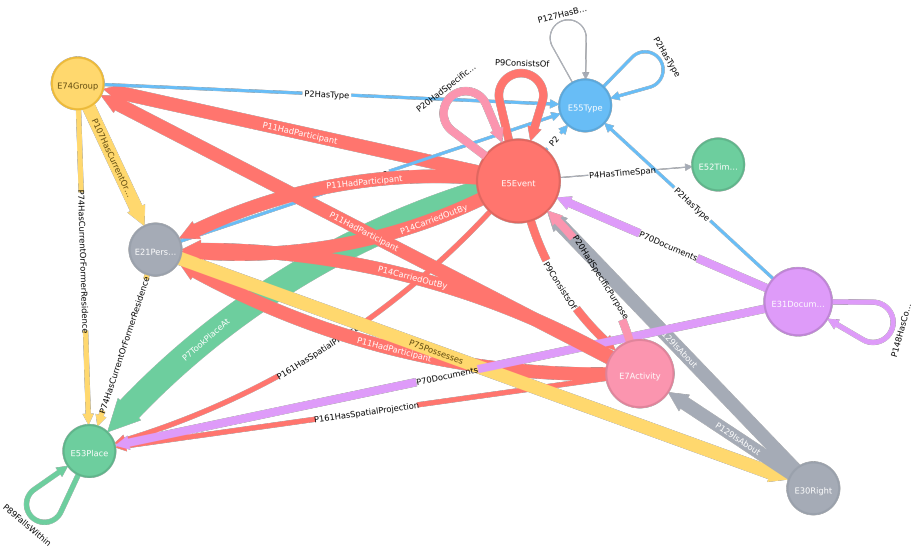

Figure 5. CIDOC-CRM data model representation. Illustration created by the authors with Neo4j Browser.

simplified data model as described in Table 6 can be illustrated as a complex graph structure (see Figure 5).

With the help of the data model implementation in the Neo4j database, custom *cypher queries*[35] can now be executed for all CRM subclasses, super classes and properties. Which charters in the dataset are type of *Bischofsurkunde* (*episcopal deeds*)? The following cypher queries produce the same results:

```
MATCH (n:E1CrmEntity)-[]-(m:E1CrmEntity) WHERE m.name="Bischofsurkunde"
RETURN n.spa_id
```

```
MATCH (n:E5Event)-[:P2HasType]-(m:E55Type) WHERE m.name="Bischofsurkunde"
RETURN n.spa_id
```

Both queries return 109 archive ids. Since all nodes are instances of the class *E1CRMEntity*[36] and only charters may have a connection to *E55Type*[37] *Bischofsurkunde*, the query results are identical. In the digital regests of the St. Katharinenspital on Monasterium.Net, up until now it has not been possible to specifically search for

35 Cypher is Neo4j's graph query language, further information in Neo4j (2018).

36 "This class comprises all things in the universe of discourse of the CIDOC Conceptual Reference Model." (Le Bœuf et al. 2015, 2).

37 "This class comprises concepts denoted by terms from thesauri and controlled vocabularies used to characterize and classify instances of CRM classes." (Le Bœuf et al. 2015, 27).

actors mentioned in a charter – if they had been annotated there – or the kind of relationship they have to a charter document.

```
MATCH (n:E5Event)-[]-(m:E21Person) WHERE m.name="Otto Prager" RETURN n.spa_id
"SpAr_Urk_86"
"SpAr_Urk_35"
"SpAr_Urk_85"
"SpAr_Urk_691"
"SpAr_Urk_692"
"SpAr_Urk_1128"
"SpAr_Urk_469"
"SpAr_Urk_76"
"SpAr_Urk_54"
"SpAr_Urk_466"
"SpAr_Urk_75"
```

The query result shows all charters in the database where *Otto Prager* is annotated as an actor. This query can now be extended, for example by an *E7Activity*, i.e., the legal form: What is the legal form of the charters in which Otto Prager is involved?

```
MATCH (a:E7Activity)-[]-(n:E5Event)-[]-(m:E21Person) WHERE m.name="Otto Prager"
    RETURN DISTINCT n.spa_id, a.name
"SpAr_Urk_86""verkaufen" (to sell)
"SpAr_Urk_85""bestätigen" (to confirm)
"SpAr_Urk_691""bestätigen" (to confirm)
"SpAr_Urk_692""verkaufen" (to sell)
"SpAr_Urk_1128""bestätigen" (to confirm)
"SpAr_Urk_1128""verpflichten" (to oblige)
"SpAr_Urk_469""schenken" (to gift)
"SpAr_Urk_76""schenken" (to gift)
"SpAr_Urk_54""verkaufen" (to sell)
"SpAr_Urk_75""überlassen" (to convey)
```

The results show that data generated by rather simple cypher queries may already provide new insights. In addition, the graph makes it possible to reveal indirect relationships between nodes. In the context of social media, this is often referred to as *Friend of a Friend* (FOAF). In our data model, the actors (*E39Actor*) are linked indirectly as witnesses or judicial actors via charters (*E5Events*) and their legal content (*E7Activity*). This interconnection represents cross-document dependencies of the St. Katharinenspital charters and is therefore particularly interesting for a closer examination. With the following query a subgraph is generated, which can be examined more closely. The Cypher query is as follows:

```
MATCH p=(n:E39Actor)-[]-(c:E5Event)-[]-(m:E39Actor)
      WHERE NOT c:E7Activity RETURN p
```

The query result consists of 662 nodes connected via 626 edges. For this quantity of nodes, a visualisation is useful. Figure 6 shows a part of the resulting subgraph. It is particularly interesting to see that there are numerous nodes that are not linked to one another outside of a very dense network. This highlights the fact that our data basis is indeed very irregular, due to the heterogeneous states of primary source analysis, and since some charter abstracts only bare very little additional information

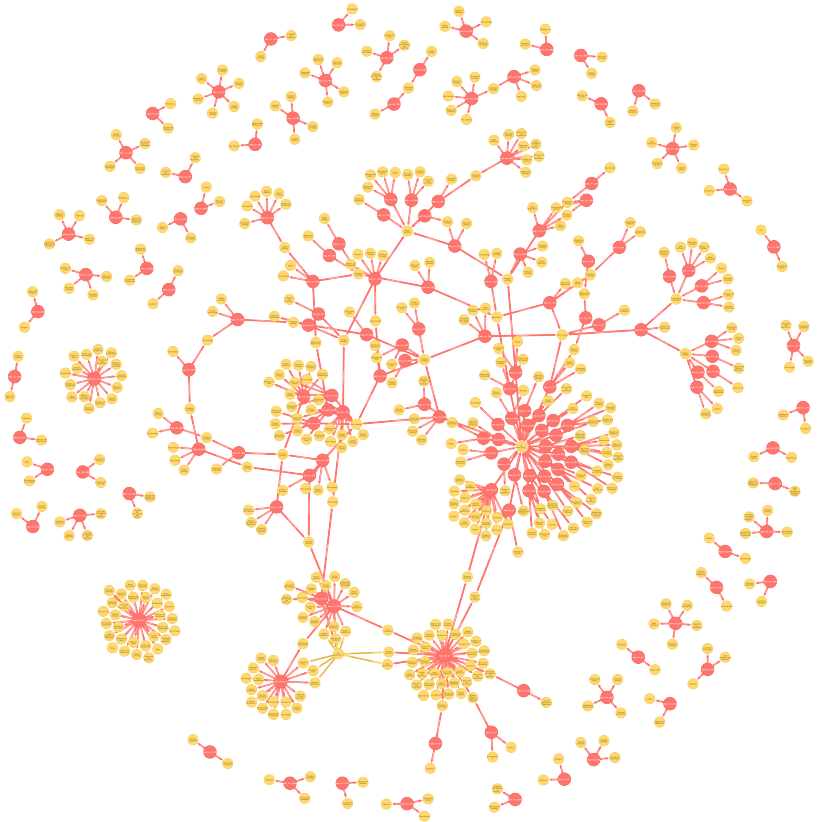

Figure 6. Linking of *E39Actors* via *E5Event* in the graph database. The network shows clearly the separation between charters not linked to any other and a very dense cluster. Illustration created by the authors with Neo4j Browser.

that can be extracted. Nevertheless, this result shows the extent to which the graph database can be used in an application that represents the various contexts of an entity, regardless of whether it is a charter, an actor or a place.

5 Creating a Web-Based Charter Portal

Charter platforms like Monasterium.Net are typically focussed on retrieving and analysing single documents. By contrast, our data model allows for further exploratory analyses (Warwick 2012, 2). We provide a web application that combines the capacities

of a full-text search[38] with queries to our graph database. Some of these queries have been discussed in the previous section. The aim of the web portal accessible at https://urkunden.ur.de is to demonstrate some the capabilities of our ontology-based data model in an application that can be used by anybody. Entities that are associated with the charters – like actors (*E39Actor*), places (*E53Place*) or documents (*E31Document*) – are shown to the user and can be accessed via persistent hyperlinks. Since the charter portal is a technical demonstration and the database is not validated yet, it should not be used like a research platform. It could, however, be used as a starting point into research related to the St. Katharinenspital charters, as it delivers a quick overview of all the data that are related to the charters. Each entity (CRM class instance) is displayed in its unique context, i.e., it is shown together with its neighbours, exactly as in the graph. We use this to enhance the result lists of a full-text search (see the coloured badges in Figure 7) and to generate entity-specific detail pages (Figure 8). The detail pages, e.g., for charters like the one shown in Figure 8,[39] show the context of the charter, which consists of related actors (*E21Person*), traditional documents, as well as other charters, that are connected indirectly over common edges to *E21Person* entities (i.e., friends of a friend). Furthermore, a hyperlink to the corresponding CEI-XML file on Monasterium.Net gives users a reference to a source that can be cited. The data model based on CIDOC-CRM also allows for the creation of class overview pages, where users get a view of all instances of a CRM class and, at the same time, traverse the CRM class hierarchy (see the yellow badge in Figure 9). Thereby, registers or entire lists of entities are created automatically via queries to the graph database.[40]

6 Conclusion

Our graph-based data model facilitates the analysis of higher-level relationships between entities and documents as it supports complex and far-reaching data queries. In our web application, users can conveniently use these queries to exploratively search the graph-based St. Katharinenspital data collection. As the graph database is traversed it delivers context-based results. The context of an entity node like a charter or an actor arises from its connection to other nodes and through its classification with the CIDOC-CRM ontology. Hence, our graph-based digital edition opens entirely new research perspectives, as the various attributes of a charter, e.g., the places of its creation, related witnesses, or traditional documents, are immediately available to

[38] For full text indexing we use Elasticsearch (Elastic 2019).

[39] CRM class *E5Event* detail page for SpAR Urk. 54 on the charter portal, available at https://urkunden.ur.de/index.php?crmentity=E5Event&prop=name&val=SpAr_Urk_54.

[40] E.g. this *E52TimeSpan* overview page lists all dated charters by their year of creation, available at https://urkunden.ur.de/index.php?crmentity=E52TimeSpan.

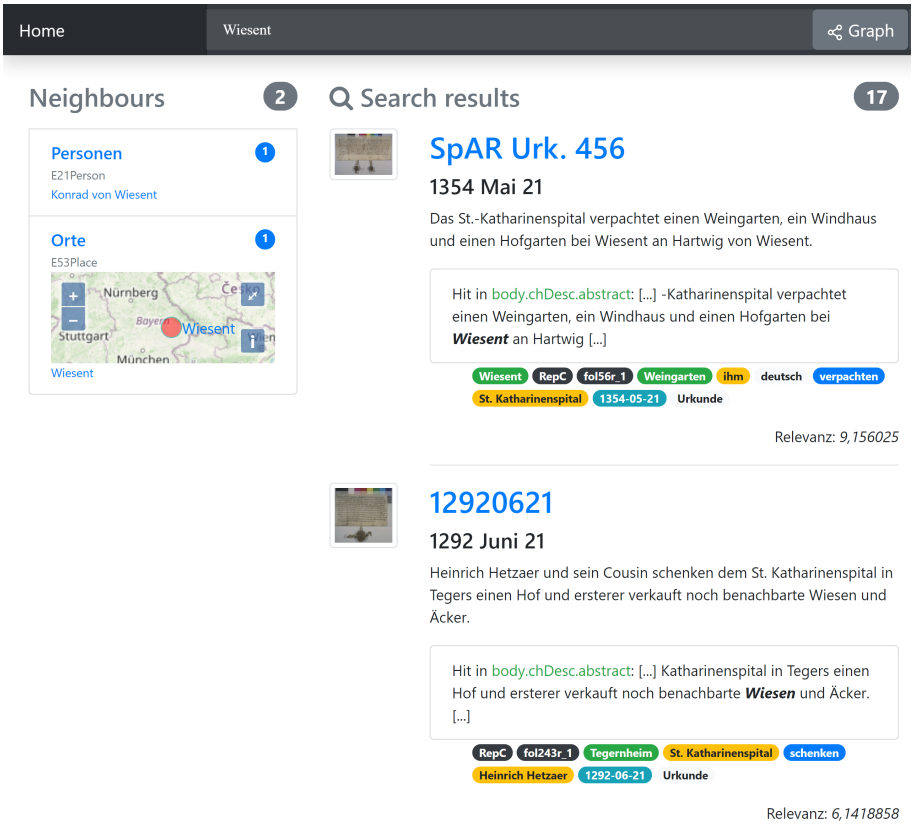

Figure 7. Full text search results for *Wiesent* (place in Bavaria). Screenshot created by the authors.

the user, jointly with their CIDOC-CRM labels. These different entities provide the facets for an experimental search portal with fulltext search capabilities. Thus, users may quickly get an overview of an entire collection of charters, and receive more information, compared to a basic fulltext search. By overcoming the limits of CEI-XML mark-up in the graph, we have created a data structure that, furthermore, facilitates the incorporation of external data sources like authority files or geospatial information. Besides this, our database can now be easily imported by charter platforms like Monasterium.Net and be used to help enhance their data basis. Further on, we showed how the application of NLP procedures facilitates the extraction of large quantities of entities from corpora of charter abstracts. We achieved this by developing a simple

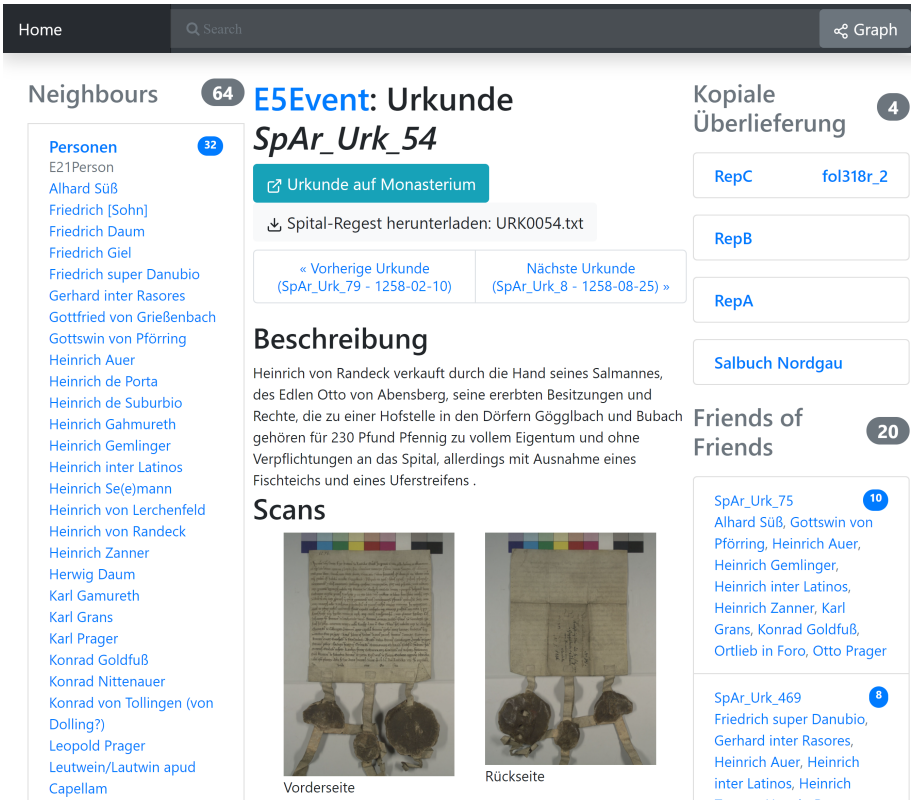

Figure 8. Graph based charter SpAR Urk. 54 and its next neighbours (CRM class *E5Event* detail page). Screenshot created by the authors.

heuristic approach. However, spaCy provides machine learning features that could provide better results, and so help to analyse more complex syntactic structures. Beyond this, our extracted data can be used to train a CRF model for improved NER in future applications. This is particularly interesting for the analysis of entire data collections e.g., on Monasterium.Net or other charter platforms. Finally, the examination of higher-level correlations by means of quantitative network analysis methods is an established approach. Therefore, a deeper analysis of graph structures is an integral part of future scenarios that could be applied to a broader dataset. Thus, the structure and the content of our graph database is to be viewed as a not-yet-completed, open-ended process.

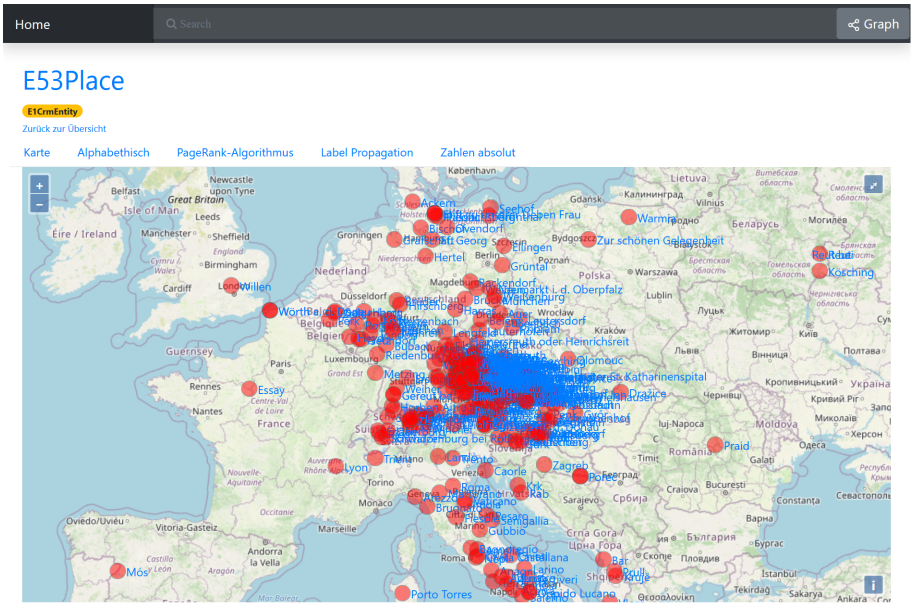

Figure 9. CRM class *E53Place* overview page[41]. Screenshot created by the authors.

Acknowledgements

We would like to thank the archive of the St. Katharinenspital in Regensburg for making the data sources available to us. Special thanks to Dr. Gernot Deinzer from the University Library of Regensburg, who provided the server infrastructure and supported our project. We would also like to thank both Dr. Žarko Vujošević from the Faculty of Philosophy of the University of Belgrade and ICARUS, who made it possible for us to present the project to an international audience at short notice.

Bibliography

Ambronn, Karl-Otto, *Verwaltung, Kanzlei und Urkundenwesen der Reichsstadt Regensburg im 13. Jahrhundert*, Münchener historische Studien (Kallmünz, Opf.: Lassleben, 1968)

Deutsche Nationalbibliothek, *Gemeinsame Normdatei (GND)*, 2018 <https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html>

41 CRM class *E53Place* overview page. Available at https://urkunden.ur.de/index.php?crmentity=E53Place.

Elastic, *Elasticsearch*, version 6.7.2 (Elasticsearch B.V., 2019) <https://www.elastic.co/de/products/elasticsearch>

Feichtmeier, Simon, *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1296 - 1301)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals et al., [preprint])

FORTH-ICS, *CIDOC CRM v6.2.1 (Draft) Encoded in RDFS* (Athen: FORTH-ICS, 2018) <http://www.cidoc-crm.org/sites/default/files/cidoc_crm_v6.2.1-2018April.rdfs>

GeoNames, *GeoNames Gazetteer*, 2018 <http://download.geonames.org/export/>

Haider, Hubert, *The Syntax of German*, Cambridge Syntax Guides, 1. publ. (Cambridge: Cambridge Univ. Press, 2010)

Honnibal, Matthew, and Ines Montani, *SpaCy - Industrial-Strength Natural Language Processing in Python*, version 2.0.16 (Berlin: ExplosionAI GmbH, 2018) <https://spacy.io/>

Jannidis, Fotis, Hubertus Kohle, and Malte Rehbein, eds., *Digital Humanities: eine Einführung* (Stuttgart: J.B. Metzler Verlag, 2017)

Jeller, Daniel, 'Urkunden als Netzwerk. Ein Werkstattbericht.', in *Quellen, Nachbarschaft, Gemeinschaft. Auf dem Weg zu einer gemeinsamen Kulturgeschichte Zentraleuropas*, ed. by Adelheid Krah (Wien; Köln; Weimar: Böhlau Verlag, 2019), 84–95 <https://dighist.hypotheses.org/945>

Kaufner, Dominik A., *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1251 - 1258)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals Regensburg, 2011)

König, Stefan, *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1145 - 1251)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals Regensburg, 2003)

Kuczera, Andreas, 'Digital Editions beyond XML – Graph-Based Digital Editions', in *HistoInformatics 2016 - The 3rd HistoInformatics Workshop. Proceedings of the 3rd HistoInformatics Workshop on Computational History (HistoInformatics 2016)*, 2016, 37–46 <http://ceur-ws.org/Vol-1632/paper_5.pdf>

———, 'Graphentechnologien in den Digitalen Geisteswissenschaften', *ABI Technik*, 37.3 (2017), 179–196 <https://doi.org/10.1515/abitech-2017-0042>

Le Bœuf, Patrick, Martin Doerr, Christian Emil Ore, and Stephen Stead, *Definition of the CIDOC Conceptual Reference Model. Version 6.2.1.*, 2015 <http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_6.2.1.pdf>

Lüschow, Andreas, 'Automatische Extraktion und semantische Modellierung der Einträge einer Bibliographie französischsprachiger Romane', in *DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts*, ed. by Christof Schöch (presented at the DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation, Parderborn, 2020), 80–84 <https://doi.org/10.5281/zenodo.3666690>

Meibauer, Jörg, Markus Steinbach, and Hans Altmann, eds., *Satztypen des Deutschen*, De Gruyter Lexikon (Berlin ; Boston: De Gruyter, 2013)

Neo4j, *Cypher Query Language Reference*, version 9, [2018] <https://s3.amazonaws.com/artifacts.opencypher.org/openCypher9.pdf>

———, *Neo4j Community Edition*, version 3.4.0 (Neo4j, Inc., 2018) <https://neo4j.com/>

Ore, Christian Emil, 'New Digital Assets - How to Integrate Them?', in *Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden*, Archiv Für Diplomatik, Schriftgeschichte, Siegel- Und Wappenkunde (Köln [u.a.]: Böhlau, 2009), XII, 238 – 254

Pasin, Michele, and John Bradley, 'Factoid-Based Prosopography and Computer Ontologies: Towards an Integrated Approach', *Literary and Linguistic Computing*, 30.1 (2015), 86–97 <https://doi.org/10.1093/llc/fqt037>

Peirson, Erick, *Cidoc-Crm-Neo4j*, version 0.1, 2017 <https://github.com/diging/cidoc-crm-neo4j>

Sahle, Patrick, *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung*, Schriften des Instituts für Dokumentologie und Editorik, 3 vols (Norderstedt: Books on Demand, 2013)

Sippl, Colin, *Charter-Abstracts*, 2019 <https://github.com/cs-ubr/charter-abstracts>

Sturm, Ferdinand, *Die älteren Urkunden des St. Katharinenspitals in Regensburg (1259 - 1270)*, Regensburger Beiträge zur Regionalgeschichte (Regensburg: Archiv des St. Katharinenspitals, 2013)

Tesnière, Lucien, *Elements of Structural Syntax* (Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2015)

Van Hage, W. R., V. Malaisé, R. Segers, L. Hollink, and G. Schreiber, 'Design and Use of the Simple Event Model', *Web Semantics: Science, Services and Agents on the World Wide Web*, 9.2 (2011), 128–136

Vogeler, Georg (ed.), *CEI - Charters Encoding Initiative* (München, 2004) <http://www.cei.lmu.de/>

———, *Mom-NLP*, 2018 <https://github.com/GVogeler/mom-NLP>

———, 'Von IIIF Zu IPIF? Ein Vorschlag für den Datenaustausch über Personen', in *DHd 2019 Digital Humanities: Multimedial & Multimodal. Konferenzabstracts*, ed. by Patrick Sahle (presented at the DHd 2019 Digital Humanities: multimedial & multimodal, Frankfurt am Main, 2019), 238–41 <https://doi.org/10.5281/zenodo.2600812>

———, 'Warum werden mittelalterliche und frühneuzeitliche Rechnungsbücher eigentlich nicht digital ediert?', *Grenzen Und Möglichkeiten Der Digital Humanities*, 1. Sonderband der Zeitschrift für digitale Geisteswissenschaften (2015), text/html Format <https://doi.org/10.17175/sb001_007>

Warwick, Claire, 'Studying Users in Digital Humanities', *Digital Humanities in Practice*, 17.5 (2012), 1–21 <https://doi.org/10.1.1.305.6694>

Appendices

Biographical Notes

Thomas Ahrend (University of Basel, Switzerland – thomas.ahrend@unibas.ch) studied Musicology, Philosophy and Literary Studies in Frankfurt a. M. and Berlin. He received his MA 1996, and his PhD 2005 at Technische Universität Berlin with a dissertation on the instrumental music of Hanns Eisler. 1997–2010 member of the editorial staff of the Hanns Eisler Gesamtausgabe in Berlin. Since September 2010, member of the editorial staff of the Anton Webern Gesamtausgabe at Musikwissenschaftliches Seminar at University of Basel.

Peter Boot (Huygens ING, The Netherlands – peter.boot@huygens.knaw.nl) studied mathematics and Dutch language and literature; he wrote his PhD thesis about annotation in scholarly digital editions and its implications for humanities scholarship. He oversaw the creation of the digital edition of the letters of Vincent van Gogh. He is employed as a senior researcher at the Huygens Institute for the History of the Netherlands where he works, among other things, as a consultant in several edition projects.

Manuel Burghardt (University of Leipzig, Germany – burghardt@informatik.uni-leipzig.de) is head of the Computational Humanities Group at Leipzig University. He is interested in the use of digital tools and computational techniques to explore new modes of doing research in the humanities. His most recent areas of research are Sentiment Analysis in the Humanities, Drametrics, Computational Intertextuality, Computational Analysis of Movies and Series and Music Information Retrieval.

Toby Burrows (University of Oxford, United Kingdom – toby.burrows@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre at the University of Oxford, and a Senior Honorary Research Fellow in the School of Humanities at the University of Western Australia.

Hugh Cayless (Duke University, USA - hugh.cayless@duke.edu) is Senior Digital Humanities Developer at the Duke Collaboratory for Classics Computing. Hugh has over a decade of software engineering expertise in both academic and industrial settings. He also holds a Ph.D. in Classics and a Master's in Information Science. He is one of the founders of the EpiDoc collaborative and currently serves on the Technical Council of the Text Encoding Initiative.

Hans Cools (University of Basel, Switzerland – 1961-2021) had a master degree in medicine and a specialization in orthopaedic surgery and traumatology (Universities of Ghent and Antwerp, Belgium, 1997), a bachelor's degree in physical

therapy, and a standalone degree in informatics (1999). Through various research and project management positions, in both companies and academic institutions, he gained expertise in different aspects of the Semantic Web technologies, focusing particularly on formal data modeling and machine reasoning. Those positions were in internationally collaborative research projects in a biomedical setting, mainly of the 5-7th EU Framework Program. Foremost in these projects were semantic interoperability and reusability of data. Since 2016, he worked in the humanities, as knowledge engineer, ontologist, and Semantic Web technology expert, at the University of Basel, as part of the NIE-INE project, which highlights scholarly editing. He (co-)published several articles, and gave workshops on the implementation of Semantic Web technologies in biomedicine and the humanities. He passed away in April 2021.

Francesca Giovannetti (University of Bologna, Italy – francesc.giovannett6@unibo.it) is a second-year PhD student in Digital Humanities at the Department of Classical Philology and Italian Studies, University of Bologna. She received an MA in Digital Humanities from King's College London and a second cycle degree in Digital Humanities and Digital Knowledge from the University of Bologna. She is interested in combining digital scholarly editing with semantic web technologies and in the use of digital technologies in education.

Matthew Holford (University of Oxford, United Kingdom – matthew.holford@bodleian.ox.ac.uk) is Tolkien Curator of Medieval Manuscripts at the Bodleian Library, University of Oxford.

Marijn Koolen (Royal Netherlands Academy of Arts and Sciences - Humanities Cluster, The Netherlands – marijn.koolen@gmail.coml) studied artificial intelligence and wrote his PhD thesis on using hyperlinks in information retrieval algorithms. He has worked on scholarly annotation for digital humanities research and on annotation-related information behaviour and information systems. He works as a researcher and developer at the Humanities Cluster of the Royal Netherlands Academy of Arts and Sciences, where he leads a project on developing annotation support within the *CLARIAH research infrastructure* project.

David Lewis (University of Oxford, United Kingdom – david.lewis@oerc.ox.ac.uk) is a Research Associate in the Oxford e-Research Centre at the University of Oxford.

Andrew Morrison (University of Oxford, United Kingdom – andrew.morrison@bodleian.ox.ac.uk) is a Software Engineer in the Bodleian Digital Library Systems and Services, Bodleian Library, University of Oxford.

Stefan Münnich (University of Basel, Switzerland – stefan.muennich@unibas.ch) studied musicology and communication science at the Technische Universität Berlin, MA 2011 with a thesis on cantional setting in Heinrich Schütz's Becker-Psalter. 2012 research assistant, 2013–2015 research associate of the Felix Mendelssohn Bartholdy. Sämtliche Briefe edition at University of Leipzig (co-editor of vols. 9 & 12). Since October 2015 research associate of the Anton Webern Gesamtausgabe, Basel; received his Doctorate degree in 2020 at the department of musicology at the University of Basel with a dissertation about music notation and its codes.

Iian Neill (Digital Academy of the Academy of Sciences and Literature, University of Mainz - Iian.Neill@adwmainz.de) is a visiting researcher at the Digital Academy of the Academy of Sciences and Literature Department at the University of Mainz, Germany. He is the creator of Codex, a text annotation environment which uses standoff property annotation to generate entities in a graph meta-model. Codex is currently being used to produce a digital edition of the epistles of Hildegard von Bingen at the Digital Academy in Mainz.

Roberta Padlina (University of Basel, Switzerland – roberta.padlina@unibas.ch) studied medieval philosophy at the University of Fribourg, Switzerland, obtaining a doctoral degree in June 2020. She has twelve years of professional experience in the field of Digital Humanities, thanks to which she has been able to work closely with different actors involved in the online publication of open access research. Roberta has worked for several years for e-codices –Virtual Library of Manuscripts in Switzerland and currently coordinates the National Infrastructure for Editions (NIE-INE) project. Roberta's main focus is on the opportunities and challenges that the digital shift poses for traditional education and research institutions, including developing semantic web strategies for scholarly publications and cultural goods.

Kevin Page (University of Oxford, United Kingdom – kevin.page@oerc.ox.ac.uk) is a Senior Researcher in the Oxford e-Research Centre and Associate Member of Faculty in the Department of Engineering in the University of Oxford.

Miller C. Prosser (University of Chicago, USA – m-prosser@uchicago.edu) earned his Ph.D. in Northwest Semitic Philology from the University of Chicago. His academic interests include the social and economic structure of Late Bronze Age Ras Shamra-Ugarit and the use of computational methods for philological and archaeological research. Miller is the Associate Director of the Digital Studies MA program at the University of Chicago where he teaches courses on Data Management and Data Publication for the Humanities. He also works as a

researcher at the OCHRE Data Service of the Oriental Institute of the University of Chicago where he consults with and supports research projects using the Online Cultural and Historical Research Environment (OCHRE). He has also worked as a tablet photographer for the Mission de Ras Shamra (Ugarit) and the Persepolis Fortification Archive Project, employing advanced digital photographic methods such as reflectance transformation imaging, photogrammetry, and high-resolution digital scanning.

Matteo Romanello (Université de Lausanne, Switzerland - matteo.romanello@unil.ch) is Ambizione SNF Lecturer at the University of Lausanne, where he conducts a project on the commentary tradition of Sophocles' Ajax. Matteo is a Classicist and a Digital Humanities specialist with expertise in various areas of the Humanities, including archaeology and history. After obtaining his PhD from King's College London, he worked as a research scientist at EPFL's DHLAB on the Linked Books and Impresso projects, before moving to his current position. He was also teaching fellow at the University of Rostock, researcher at the German Archaeological Institute, and visiting research scholar at Tufts University.

Sandra Schloen (University of Chicago, USA – sschloen@uchicago.edu) is the Manager of the OCHRE Data Service at the Oriental Institute of the University of Chicago, and is the co-designer and developer of the Online Cultural and Historical Research Environment (OCHRE). Trained in computer science and mathematics (B.Sc. University of Toronto; M.Ed. Harvard University), Sandra has spent over 30 years working with technology as a systems analyst, technical trainer, and software developer. A long association with colleagues in the academic community has enabled her to develop a specialty in solving problems in the Digital Humanities where challenges of data capture, data representation and data management abound. Specifically, she has served extensively as a database manager for several archaeological projects in Israel and Turkey, and supports a wide range of research projects at the Oriental Institute and at other universities.

Desmond Schmidt (University of Bologna - desmond.allan.schmidt@gmail.com) has a background in classical Greek philology, information security and eResearch. He has worked on several scholarly edition projects, including the Vienna Wittgenstein Edition (1990–2001), Digital Variants (2004–2008), the Australian Electronic Scholarly Editions project (2012–2013), the Charles Harpur Critical Archive (2014-) and a pilot edition of Gianfrano Leopardi's Idilli (2018-). He currently works on developing practical web-based tools for making, visualising and publishing digital scholarly editions.

Colin Sippl (University of Regensburg, Germany – colin.sippl@ur.de) is currently a project employee at the University Library of Regensburg. Since 2017, he has been working on extending the open access services of the Electronic Journals Library (EZB). More recently, he has started developing and setting up a digital repository for literature, artefacts and experiments relating to the early life sciences based on the Invenio framework. He specialised in textual data mining and the development of media services in the institutional domain.

Elena Spadini (University of Lausanne - elena.spadini@unil.ch) is a postdoctoral researcher at the University of Lausanne. She holds a Ph.D. in Romance Philology from the University of Rome Sapienza (2016) and a M.A. in Digital Humanities from the École nationale des chartes (2014). She was a Marie Curie fellow in the IT Network DiXiT and co-directed the related volume Advances in Digital Scholarly Editing (Sidestone Press, 2017). She published in international journals and taught specialized courses in various European countries in the field of Digital Philology.

Francesca Tomasi (University of Bologna - francesca.tomasi@unibo.it) is associate professor in Archival Science, Bibliography and Librarianship at the University of Bologna (Italy). Her research is mostly devoted to digital cultural heritage, with a special attention to documentary digital edition, and a focus on knowledge organization methods in archives and libraries. She is member of different scientific committees of both associations and journals. In particular, she is President of the Library of the School of Humanities in the University of Bologna (BDU - Biblioteca di Discipline Umanistiche), Director of the international second cycle degree in Digital Humanities and Digital Knowledge (DHDK), President of the Italian Association of Digital Humanities (AIUCD – Associazione per l'Informatica Umanistica e la Cultura Digitale), and co-head of the Digital Humanities Advanced Research Center (/DH.ARC). She wrote about 100 papers and 4 monographs related to DH topics. She is editor and scientific director of several digital scholarly environments.

Athanasios Velios (University of the Arts London, United Kingdom – a.velios@arts.ac.uk) is Reader in Documentation at the University of the Arts London.

Georg Vogeler (University of Graz - georg.vogeler@uni-graz.at) is professor for Digital Humanities at the University of Graz and scientific director of the Austrian Center for Digital Humanities and Cultural Heritage at the Austrian Academy of Sciences. He is a trained historian (Historical Auxiliary Sciences). He spent several years in Italy (Lecce, Venice). In 2011, he became member of faculty at the Centre for Information Modelling at Graz University, where he was nominated

full professor for Digital Humanities in 2016 and head of department in 2019. His research interests lie in late medieval and early modern administrative records, diplomatics (digital and non digital), digital scholarly editing and the history of Frederic II of Hohenstaufen (1194–1250). He was and is part in several national and international research projects related to his research interests.

Christian Wolff (University of Regensburg, Germany – christian.wolff@ur.de) has been Professor of Media Informatics at the Institute for Information and Media, Language and Culture at the University of Regensburg since 2003. He holds a PhD in information science and is a habilitated computer scientist. His research interests include: human-computer interaction, multimedia and web-based information systems, (multimedia) software engineering and information retrieval (in particular information literacy and social media).