Semantically Enriched Historical Data. Drawing on the Example of the Digital Edition of the "Urfehdebücher der Stadt Basel"

Christopher Pollin and Georg Vogeler[1]

University of Graz, Centre for Information Modelling, Austria
<christopher.pollin@uni-graz.at>, <georg.vogeler@uni-graz.at>

Abstract. Historical data is widely recognized as a rather complex type of data that contains records about multi-layered, context-sensitive entities and can often be represented as a graph. This paper describes the digital edition of the "Urfehdebücher der Stadt Basel" as an example of how semantic web technologies can offer comprehensive tools in response to the challenges coming with historical data. It introduces the FEDORA Commons based GAMS-infrastructure, reports the workflow from XML/TEI[1] encoded historical documents to semantically enriched data in form of XML/RDF data, and describes the specific data model for the resource. Finally, the paper discusses how the data can be used beyond a standard web interface with reading and search functionalities, for analysis with network visualisation functionalities.

Keywords: GAMS, historical data, digital edition, semantic enrichment, Urfehde, TEI, RDF, SKOS, data visualisation

1 Introduction

The High Level Expert Group on Scientific Data formulated their shared vision for 2030: '*Our vision is a scientific e-infrastructure that supports seamless access, use, re-use and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure.*' [Neuroth et al. 2012]
Scientific data is always related to the context of a scientific problem. Research data in the humanities, including historical data, is interlinked to its scientific discipline and tends to be complex in a specific way. [Thaller 1989] points out particular challenges concerning historical data: Historical terms, for example 'Prussia', can vary in relation to spatial and temporal context. This leads to a definition of historical data by [Meroño-Peñuela / Hoekstra 2014] as the union of a static, unique primary source and dynamic secondary sources, where the latter point at the primary source in different time- and context-sensitive ways. For this reason the authors recommend to describe historical data as a graph and connect it to linked open data sources using taxonomies or ontologies on the

[1] http://www.tei-c.org, 21.7.2017

one hand, and dereferencing services inside a digital archive on the other hand. Thinking of the mentioned vision and the fact of historical data being multi-layered and context-sensitive, semantic web technologies can offer comprehensive tools that address these problems. The aim of this paper is it to outline the process of semantical enrichment of a historical dataset, the '*Urfehdebücher der Stadt Basel*', as the primary source, and their representation as secondary sources, the '*Urfehdebücher der Stadt Basel – digitale Edition*' [Burghartz / Calvi / Vogeler 2016]. The process of semantically enriching and formalizing data using semantic web technologies could fulfil the vision of data becoming its own infrastructure.

The text of the digital edition of the 'Urfehdebücher der Stadt Basel – digitale Edition' was created in a small scale project by Susanna Burghartz at the University of Basel in a teaching project together with her students, with particular contributions by Sonia Calvi and Anna Reiman. The technical realization was developed by the Centre for Information Modelling at the University of Graz. The aim of this low-budget and student supported project lies more in an experimental approach applying semantic web technologies to a historical source. 'Urfehde' can be roughly translated as 'oath of truce'. The purpose of the so called 'Urfehde' was to settle a dispute between two conflict parties and urge a sentenced criminal to a unilateral oath not taking revenge for its judge. This was legal practice in most of central Europe in the late middle ages and the early modern period, and recorded in the so called 'Urfehdebücher', which have survived in many archives, as demonstrated in the data of the *Index Librorum Civitatum* project.[2] The first *Urfehdebuch X* of the city of Basel (StadtA Basel Ratsbücher O10) records 'Urfehde' oaths from 1563 to 1569. This source can be used as an exemplary dataset as it shows the significant structure to be used in a statistical analysis: in addition, with 625 entries, the dataset is large enough to contribute to research on the cultural, social, and economic history of early modern people.

The task in realising this digital edition was therefore to combine established and easy to use transcription workflows using XML/TEI annotation with the conversion to RDF data to prepare a basis for data analysis and data publication. This calls into question which advantages semantic web technologies can offer to scholars regarding the retrieval, visualisation and analysis of historical data in the humanities.

1.1 Related Work

Maybe the first digital edition making use of semantic web technologies in a similar way was the edition of the Henry III fine rolls[3] [Ciula et al. 2008]. The project combined a TEI transcription with a CIDOC-CRM based ontology expressed in OWL. The RDF data of the project was not made openly accessible. The *Henry III fine rolls* project follows the approach of building an *extended*

[2] http://www.stadtbuecher.de/literatur/schlagwort/137667, 18.07.2017.
[3] http://www.finerollshenry3.org.uk, 12.7.2017

index for the digital representation of the primary source as [Poupeau 2006] has described it. A successful example for this approach is '*Sandrart.net*'. This digital scholarly edition encodes data using XML/TEI. The data is made available Linked Open Data in RDF.[4] Similar to this the platform for historical research SYMOGIH[5] is preparing a SPARQL endpoint. Recently the '*Semantic Blumenbach*' project explores new approaches for linking between artefacts and text [Wettlaufer et al. 2015]. It uses the 'scientific communication infrastructure' *WissKI*[6] to implement semantic web methods for data acquisition, storage and re-use.

All these projects follow the *extended index* approach. This, however, fails short when it comes to the analysis of abstract concepts apart from the classical index on named entities like places, persons, and objects. Historical texts like the *Urfehdebücher* need additional modelling to become data sets for historical analysis, in particular regarding the classification of criminal offence, punishment, and social status of the people involved. Thus the '*The Proceedings of the Old Bailey*' project comes much closer to the *Urfehdebücher*. The project defines itself as a searchable edition of criminal trials held at London's central criminal court. XML/TEI markup of digitized text offers the possibility to search and analyse the source.[7] The data set is accessible via an API[8], but is not available as RDF or via a SPARQL endpoint. Therefore there is no project in the same research area as the 'Urfehdebücher'. Common standards have still to be established and the challenges of interoperability are not solved yet.

2 The digital Edition

The workflow of the *Urfehdebücher*-project is embedded in the GAMS[9], which is described by [Steiner / Stigler 2017]. GAMS defines itself as an asset management system for the humanities and serves the purpose of administration, publication and long-term preservation of digital resources. It is based on the open source repository software FEDORA-Commons. Using Cocoon-services and project specific content models for varying data streams scholarly data can be stored and disseminated for public use. The data is represented as readable web site, as archival data structures in XML, and via various API. GAMS implements a disseminator for RDF data via a RDF-triplestore. Currently, the open source software Blazegraph[10] is in use, which allows SPARQL-queries and full-text search in literals.

Expert academics transcribed and encoded the source in XML/TEI, structuring the text, marking up text-specific phenomena and normalizing places, persons

[4] http://ta.sandrart.net/de, 12.07.2017
[5] http://symogih.org, 12.7.2017
[6] http://www.wiss-ki.eu/, 12.7.2017
[7] https://www.oldbaileyonline.org, 12.07.2017
[8] https://www.oldbaileyonline.org/static/API.jsp, 12.7.2017
[9] gams.uni-graz.at, 12.7.2017
[10] https://www.blazegraph.com, 12.07.2017

or concepts. Additionally the TEI attribute `ana` was used to add specific semantics to the applied TEI markup. `ana` is used because it allows to add global and multiple interpretation to the TEI markup. This XML/TEI illustrates the TEI markup.[11] The `div` element with the attribute `ana="#uf_Eintrag"` defines the content of the whole `div` as an 'Urfehde'-entry representing a single case. The semantics of the first entry in the XML/TEI can be summarized as follows: The hireling '*Heinrich Peter*' from '*Zurich*' was judged as an offender, due to alcohol abuse on the '*kornmerkt*' (grain market). This statement is encoded in the XML/TEI using the attribute `ana`, like `ana="#uf_male"` for annotating the gender of a person. The value in the `ana` attribute is taken from a taxonomy of categories defined by the the colleagues in Basel following their methodological access of the source[12]. Its hierarchical structure of concepts can easily be converted into a SKOS-resource. When ingesting the XML/TEI into the GAMS infrastructure, a project-specific XSLT-Stylsheet transforms all semantically enriched data into XML/RDF and writes the triples in the triplestore. The XML/RDF shows the outcome of the transformation.[13] The assertions describe the aforementioned 'Urfehde'-entry and all its properties linking to other concepts like `uf:PersonOffender`, where further properties refer to literals or refer to concepts normalizing data, like the place '*Zürich*'.
The extracted XML/RDF follows a simple RDFs[14] which defines the entry (`Eintrag`) at the core. It represents the case and has properties to identify the offence and its classification, the persons named in the record and their role, the type of punishment, and other properties connected directly to the entry and the legal procedure, e.g. date of the oath (`DatumUrfehde`), date of the offence (`DatumTat`), the notarial authentication of the entry (`NotarialSubscription`). Fulltext (advanced) search functionalities are implemented using SPARQL and the fulltext capabilities of the *blazegraph* triple store.[15] For this purpose GAMS offers a query content model which returns XML data on demand. This can subsequently be transformed to HTML to offer additional functionalities like visualisations and data download.

3 Results and Potentials

The outcome is a semantically enriched digital edition using RDF data representation[16] for fulltext and advanced search.[17] Using the advanced search functionalities a user is able to employ regular expressions, make temporal constraints of search results and use the normalization of place names to query the data.

[11] gams.uni-graz.at/o:ufbas.1563/TEI_SOURCE, 12.09.2017
[12] gams.uni-graz.at/o:ufbas.kategorien/TEI_SOURCE, 12.07.2017.
[13] gams.uni-graz.at/o:ufbas.1563/RDF, 12.09.2017
[14] gams.uni-graz.at/o:ufbas.schema, 15.7.2017
[15] wiki.blazegraph.com/wiki/index.php/FullTextSearch, 12.7.2017
[16] gams.uni-graz.at/o:ufbas.1563/RDF, 13.07.2017.
[17] gams.uni-graz.at/query:ufbas.search/get, 13.07.2017.

Regular expressions are particularly useful to so search for orthographic alternatives, e.g. *eefrou?wen* which returns words like *eefrouwen* and *eefrowen* (for 'spouse'). An example for using normalized data is that a query like *Kleinbasel* returns all entries connected to the place named *mindren Basell* in the text.
A navigation menu leads through the chronologically listed data. The user can collect entries from search results or while browsing the text into a personal data basket[18], implemented by using the local storage of the browser and simple JavaScript. Collected entries can be exported as simple CSV to be processed with a spreadsheet application for further work.
Because of the fact that the whole data set is defined as RDF graph and the data itself has network character, adequate ways of information visualization are possible. Exemplary scholarly questions regarding the 'Urfehdebücher' could be if female offenders of a specific type of crime were treated and punished differently than male offenders. Visualizing the relations between offender, places, time, punishment or crime in the whole data set, or parts of it, could open new approaches to work with the source, or open possibilities to identify at a glance which category or question could be interesting. We did some experiments using d3.js[19] library for creating forced graphs, based on the result of the search for a category.
This Figure shows a graph of the search by category `uf:ThreatOfPunishment`.[20] The light green node in the center represents this category. Every dark blue node refers to a case reported in the 'Urfehdebuch', which is connected to the node with the value `uf:male` (large blue node). The light blue nodes represent cases connected to women (`uf:female`, large yellow node). The other paths from the case nodes represent dates (light blue), occupations (green), and places of origin (orange). The gender nodes are obviously the major bridge-nodes, but other properties to the cases can serve as additional bridge-nodes, e.g. when several cases contain the same date (*1568-06-14*), same profession (*taglöner*) or same place (*Zürich*). The forced atlas graph allow a first instantaneous interpretation: The degree and centrality of the gender nodes moves the node for female offenders at the outer part of the graph. A threat of punishment was therefore much more often applied to male offenders, and alcohol abuse was a problem recorded mostly for men. Certainly detailed research has to establish the numbers relative to the number of all cases. The graph visualization can assist retrieval and discover functionalities in the future.

4 Conclusion and Further Work

The example of the *Urfehdebücher* demonstrates that creating XML/TEI transcription of a text prepared to be used as semantic web data offers new approaches for scholarly edition, fits to the graph-like understanding of historical

[18] gams.uni-graz.at/context:ufbas?mode=datenkorb, 13.07.2017.
[19] https://d3js.org, 21.07.2017.
[20] gams.uni-graz.at/context:ufbas/StrafeStrafandrohung, 12.09.2017

data, and the data becomes more expressive and self-describing. The transformation of the textual statements in RDF is made with easy annotation and little programming effort. The RDF dataset can be used as a fundamental database technology in the online publication as well as for advanced research questions. The data created can be queried and visualized in a way that it can be beneficial for historical research. Finally the publication of this data with semantic web technologies allows to make the data model, the taxonomy and the data itself openly available in a standardized way as RDFs, SKOS and generic RDF data. Aligning the data model and the taxonomy with other resources like the *Old Bailey* project is envisioned future improvement and can be the first step to a common vocabulary.

5 Appendix

- TEI-Source: gams.uni-graz.at/o:ufbas.1563/TEI_SOURCE
- RDF-Source: gams.uni-graz.at/o:ufbas.1563/RDF
- Graph of `uf:ThreatOfPunishment`: gams.uni-graz.at/context:ufbas/StrafeStrafandrohung

References

[Burghartz / Calvi / Vogeler 2016] Burghartz, Susanna / Calvi, Sonia / Vogeler, Georg: Urfehdebücher der Stadt Basel – digitale Edition, Graz 2016, gams.uni-graz.at/ufbas.

[Ciula et al. 2008] Ciula, Ariana / Spence, Paul / Veira, José Miguel: Expressing complex associations in medieval historical documents. The Henry III Fine Rolls Project, in: *Literary and Linguistic Computing* 23 (2008), p.311–325, DOI: 10.1093/llc/fqn018.

[Meroño-Peñuela / Hoekstra 2014] Meroño-Peñuela, Albert / Hoekstra, Rinke: What is linked historical data?, in: *International Conference on Knowledge Engineering and Knowledge Management*. Springer, Cham, p.282–287.

[Neuroth et al. 2012] Neuroth, Heike / et al.: *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*. Hülsbusch, 2012, p.15.

[Poupeau 2006] Poupeau, Gautier: De l'index nominum à l'ontologie. Comment mettre en lumière les réseaux sociaux dans les corpus historiques numériques?, in: *Digital Humanities 2006. The First ADHO International Conference: Conference Abstracts. Université Paris-Sorbonne*. 2006.

[Steiner / Stigler 2017] Steiner, Elisabeth / Stigler, Johannes : *GAMS and Cirilo Client. Policies, documentation and tutorial*. Graz, 2014–2017 http://gams.uni-graz.at/docs.

[Thaller 1989] Thaller, Manfred: The Need for a Theory of Historical Computing, in: Denley, Peter / et al.: *History and Computing II*, Manchester and New York, 1989, p.4–6.

[Wettlaufer et al. 2015] Wettlaufer, Jörg / et al.: Semantic Blumenbach. Exploration of Text–Object Relationships with Semantic Web Technology in the History of Science, in: DSH *Digital Scholarship in the Humanities* 30, suppl. 1. 2015, p.187–198.