Originally published in: Digital Humanities 2014. Conference Abstracts, EPFL - UNIL, Lausanne, Switzerland, 8-12 July 2014. - Lausanne: EPFL/UNIL, 2014. Pp. 95-98.

Uncertain about Uncertainty: Different ways of processing fuzziness in digital humanities data

Binder, Frank
Universität Gießen, Germany

Entrup, Bastian
Universität Gießen, Germany

Schiller, Ines
Universität Gießen, Germany

Lobin, Henning
Universität Gießen, Germany

1 Introduction

The GeoBib project is constructing a georeferenced online bibliography of early Holocaust and camp literature published between 1933 and 1949 (Entrup et al. 2013a). Our immediate objectives include identifying the texts of interest in the first place, composing abstracts for them, researching their history, and annotating relevant places and times. Relations between persons, texts, and places will be visualized using digital maps and GIS software as an integral part of the resulting GeoBib information portal.

The combination of diverse data from varying sources not only enriches our knowledge of these otherwise mostly forgotten texts; it also confronts us with vague, uncertain or even conflicting information. This situation yields challenges for all researchers involved – historians, literary scholars, geographers and computer scientists alike. While the project operates at the intersection of historical and literary studies, the involved computer scientists are in charge of providing a working environment (Entrup et al. 2013b) and processing the collected information in a way that is formalized yet capable of dealing with inevitable vagueness, uncertainty and contradictions. In this paper we focus on the problems and opportunities of encoding and processing fuzzy data.

2 The uncertainty about uncertainty: How to model and represent it

The data collected in our project concerns such different entities as texts, persons and places and is compiled from different sources and different scholarly perspectives. The project is entirely interdisciplinary: besides literary scholars and historians, also geographers and computer scientists are involved. Students and researchers from literary and historical studies are our target audience for whom the resulting online platform shall provide an attractive research tool. Hence, the collected data does not only lie in the intersection of research interests of these fields, but extends to the sum of these interests. The platform is supposed to help answer questions that arise in the field of literature, e.g. finding texts concerned with certain places in a given time period, but also to support historians in finding possible eye witness reports of the crimes of Nazi Germany. Accordingly, information of various kinds is collected with the intention of supporting such diverse use cases. The different scholarly perspectives also determine the amount and kind of data we collect, and their information needs can hardly be covered by a single formalism or predefined ontology. We need a flexible yet coherent formalization that is adaptable to our objectives.

Our workflow and approach to collecting data is one of *divide et impera*: Instead of proposing one format that does it all, we distinguish between different kinds of information depending on the entities concerned. Information collected on the authors of the holocaust texts and relevant places is stored in a user friendly MediaWiki system, while information on these texts is stored in TEI/XML files. Both systems are interconnected and geographical references are integrated as well (cf. Entrup et al. 2013b). The resulting information portal will be backed by an object-relational database.

2.1 Persons and places: Capturing FUZZY information in a Wiki System

Within the field of prosopography the combination of different, possibly contradicting sources is a well-known problem. Pasin and Bradley (2013) offered insights on how such alternative views on historical events could be described using an underlying ontology. Software libraries intended to support processing of prosopographic data are also being developed (e.g. Barabucci and Zingoni, 2013). The GeoBib project collects information and short biographies of authors – a task that bears resemblance to prosopographic research. Many of those authors only published one text. Researching their personal information often leads to ambiguous results, such as different names used, differing information on birth or death dates as well as other personal data.

We extended the MediaWiki system that we use to collect information on persons and places with a set of templates that help to ensure that such information is added in a coherent way, while allowing the data to be vague or apparently contradictory. The Wiki allows the editors to add uncertain information into proposed fields, so that, for instance, different names can be added to one person. Furthermore, the short biographical texts we compose for most of the authors can be used to communicate dissent between different sources.

2.2 TEI/XML: Encoding uncertainty in literary annotations

Literary texts, and as such especially autobiographies and memoires, are not collections of historical facts arranged in an exact chronology. Especially the early Holocaust texts are emotionalizing (cf. Feuchert, 2012, Hickethier 1986) and "conveying the experience made with the National Socialist terror system in a literary, or better, literarised way" (Feuchert, 2012, p. 218). But even apparently factual accounts of events carry a certain degree of vagueness, which leads both historians and literary scholars to interesting research questions, and poses a challenge for data modeling and database design.

Vagueness already occurs when collecting formal metadata. For each holocaust text under consideration we try to identify the first published edition. Yet looking at some more widely known texts of that era, we find multiple editions that differ in such basic information as publisher, year of publication, editor, or even the title. Such phenomena are familiar to those concerned with bibliographic information. Special care is required when formalizing such inconsistent data. In our collection of TEI/XML files, every single edition is represented by one XML file. These files contain the bibliographic information in the TEI header, and they are linked to the corresponding other editions.

TEI provides the @cert attribute for indicating (un)certainty of information given in an XML node. While this strategy allows us to indicate possible vagueness in a machine-readable way, we also need to find ways of communicating this uncertainty to the human user of our information portal. For that purpose, uncertainty regarding the plot or the history of reception of a literary work is captured in literary annotations, i.e. running text, which is an effective and straight-forward way of communicating uncertainty to human readers. In this context, we also allow adding <note> elements to be used by our editors for supplying small texts that will be presented to the user describing the kind of uncertainty involved (cf. Bradley and Short, 2005). The combined use of both elements allows conveying uncertainty to the human user while keeping it encoded in a machine-readable (or machine-traceable) way.

2.3 Modeling Uncertainty in a Database

The GeoBib information portal will rest on a PostgreSQL database (Scherbaum 2009) [4], where we use object relations for the description of certainty and note elements. Our first example represents a person (see Figure 1a). There are three different names associated with the entity: a birth name, the name after her wedding, and a pseudonym.

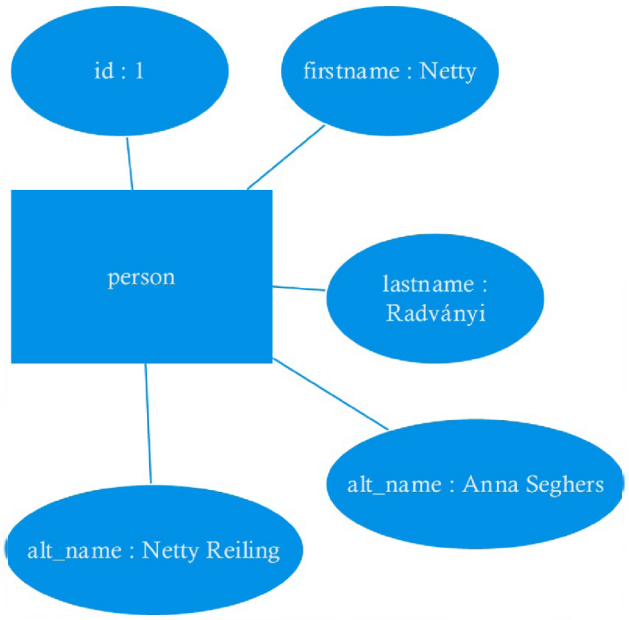

Fig. 1: Exemplary database entries for a person : a) person entity model

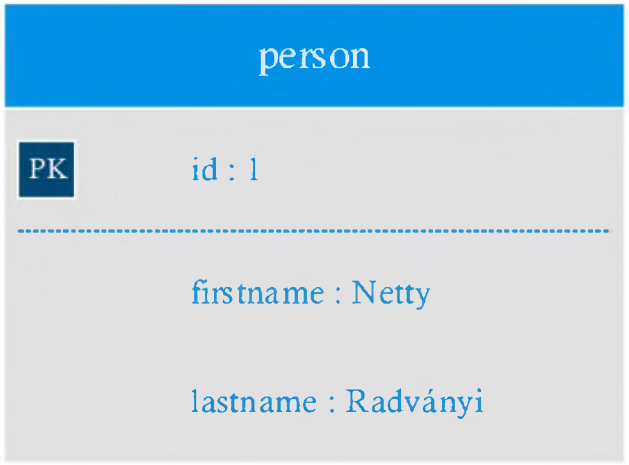

Fig. 2: Exemplary database entries for a person : b) simplified perso object

Fig. 3: Exemplary database entries for a person : c) related alternative/ additional information

While a person entity has certain fields that can be filled in (see Fig. 1b), we use object relations (Fig. 1c) to add alternative information and related values of certainty and/or a note.

The second example describes a literary work with an uncertain year of publication. A relational database would require intermediate tables for all attributes of one entity, which may have uncertainties and/or notes attached (cf. Bradley and Short, 2005). In an object relational database using one special entity is sufficient in such a case.

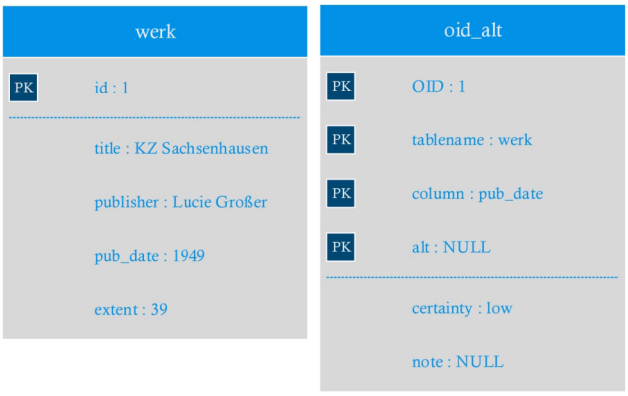

Fig. 4: Work entity and related certainty field "pub_date"

As shown in Fig. 2, the table includes the object ID, the table name, the relevant column name, and the alternative content. Each dataset may contain a certainty attribute and/or a note. The certainty field is defined, in accordance with TEI, as either {high, medium, low, unknown} and the note field is a text that will be presented to the user and is meant to explain the uncertainty[6]. In the example above (Fig. 2) a year of publication is given but its certainty is marked as "low". Such relations can be added for every field of every entity in the database.

3 Discussion and Outlook: Surely more uncertainty

We have just presented our approach of encoding uncertainty in our database. Such information can be used, for instance, to rank search results or to increase recall on certain queries and parameter settings. But we still see more challenges ahead: The encoded uncertainty has to be communicated effectively to the human user. Accordingly, the visualization of uncertainty, and especially the presentation of search results based on uncertain or divergent information will be among our next concerns. A similar problem of visualizing uncertainty arises in the forthcoming georeferencing and geotagging: Literary texts are no geographical maps. They constitute themselves in their geographical space but might encode this information in a way hard to decipher (cf. Reuschel et al., 2013). Fictional place names can sometimes be identified with actual places on a map, but sometimes it is impossible to do so. Geographical locations may be referred to by metaphors, old or forgotten names, local identifiers or nicknames. Such informal references frequently remain geographically imprecise and require interpretation (cf. Hill, 2006, p. 28f). The textual material that our editors provide could also be used to test and improve automatic methods of geographical relation extraction (e.g. Blessing and Schütze, 2010). Still, automatic georesolving is hindered by the limited (historical) coverage of contemporary gazetteers, spelling changes and changing administrative boundaries (Tobin et al., 2010, p. 8). Such limitations also pertain to prisons and camps, those places of high interest in our domain, whose exact geographical locations have to be reconstructed manually before adding them to our databases.

4 Acknowledgements

Funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) from July 2012 to June 2015 (FKZ: 01UG1238A-B).

References

Barabucci, Gioele and Jacopo Zingoni (2013). *PROSO: prosopographic records. In: Proceedings of the 1st Workshop on Collaborative Annotations* in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities, DH CASE '13. September 10 2013, Florence, Italy <doi:10.1145/2517978.2517982>

Blessing, Andre and Hinrich Schütze (2010). *Self-annotation for fine-grained geospatial relation extraction.* In: Proceedings of the 23rd International Conference on Computational Linguistics pp. 80-88. dl.acm.org/citation.cfm?id=1873781.1873791

Bradley, John, and Harold Short (2005). *Texts into Databases: The Evolving Field of New-Style Prosopography.* In: Literary and Linguistic Computing, 20 (2005), 3–24 <doi:10.1093/llc/fqi022>

Entrup, Bastian, Maja Bärenfänger, Frank Binder and Henning Lobin (2013a): *Introducing GeoBib: An Annotated and Geo-referenced Online Bibliography of Early German and Polish Holocaust and Camp Literature (1933–1949)*. Digital Humanities 2013, University of Nebraska–Lincoln, 16-19 July 2013. dh2013.unl.edu/abstracts/ab-229.html

Entrup, Bastian, Frank Binder and Henning Lobin (2013b): *Extending the possibilities for collaborative work with TEI/XML through the usage of a wiki-system*. In: Proceedings of the 1st Workshop on Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities, DH CASE '13. September 10 2013, Florence, Italy. <doi:10.1145/2517978.2517988>.

Feuchert, Sascha (2012). *Fundstücke: Bemerkungen zu Darstellungskonventionen und paratextuellen Präsentationsformen früher Texte deutschsprachiger Holocaustliteratur.* In: Günter Butzer / Joachim Jacob (Hg.): Berührungen. Komparatistische Perspektiven auf die frühe deutsche Nachkriegsliteratur. München: Wilhelm Fink 2012, pp. 217-230.

Hickethier, Knut (2006). *Biographie*, *Autobiographie*, *Memoirenliteratur.* In Ludwig Fischer (eds), Literatur in der Bundesrepublik bis 1967. München 1986, pp. 574–584.

Hill, Linda L. (2006). *Georeferencing*. The Geographic Associations of Information, Cambridge: The MIT Press.

Pasin, Michele, John Bradley (2013). *Factoid-based Prosopography and Computer Ontologies: towards an integrated approach.* In: Literary and Linguistic Computing (2013). <doi:10.1093/llc/fqt037>

Reuschel, Anne-Kathrin, Barbara Piatti and Lorenz Hurni (2013). *Modelling Uncertain Geodata for the Literary Atlas of Europe.* In: K. Kritz et al. (eds.) Understanding Different Geographies. Lecture Notes in Geoinformation and Cartography, <doi:10.1007/978-3-642-29770-0_11>

Scherbaum, Andreas (2009). P*ostgreSQL – Datanbankpraxis für Anwender, Administratoren und Entwickler.* Open Source Press. München, 2009.

Tobin, Richard, Claire Grover, Kate Byrne, James Reid and Jo Walsh (2010). *Evaluation of georeferencing.* In: Proceedings of the 6th Workshop on Geographic Information Retrieval, GIR '10, Zurich, Switzerland, <doi:10.1145/1722080.1722089>

As of December 2013 the GeoBib project has identified and collected bibliographical information on 670 texts of early Holocaust and camp literature and produced 130 annotation documents so far. These include references to 620 authors, 550 locations, 230 camps and 45 ghettos.

See for instance svario.it/factoid

German original: „*Bereits seit 1933 sind, vor allem im Ausland, Texte erschienen, die Erfahrungen mit dem nationalsozialistischen Terrorsystem literarisch oder besser: literarisiert vermitteln*“ (Feuchert, 2012, p. 218)

http://www.postgresql.org/about/

We do not use the PostgreSQL entity parameter “WITH OIDS”, because of the high memory requirements (cf. Scherbaum 2009, p. 161f). Our OID (object identifiers) are composed of the entity id and name.

Alternatively, the note field could also contain additional information.