Modeling Patterns in Written Natural Language Questions to Archives

Steffen Hennicke
Humboldt-Universität zu Berlin
Berlin School of Library and Information Science
Germany
steffen.hennicke@ibi.hu-berlin.de

Abstract

This short paper is part of an ongoing dissertation project and introduces the idea to create an ontological model – the *Archival Knowledge Model* (AKM) – of common patterns found in written natural language questions to archives. Such an ontological model can be used to analyze and query archival knowledge bases in order to provide more adequate answers and to enable more relevant discovery facilities. For this purpose, written reference questions to the German Federal Archive, the *Bundesarchiv*, are being analyzed and patterns found translated to the CIDOC CRM and appropriate extensions.

1 Introduction

Archives hold enormous *information potential* [MH01] which are meant to be explored and accessed through archival aids as well as the expertise of archivists. Although the conceptualization of these descriptive tools is based on elaborate and historically grown archival principles and models, their design is less informed by explicit knowledge about the information needs of archival users [Cox08]. Digital representations of these archival aids typically emulate the original descriptive structures and render a vast amount of information implicit. At the same time, search facilities are mostly simple search interfaces which only allow keyword based searches and return plain lists of matches.

Research shows that such search and retrieval systems do not properly serve the users. One of the pivotal reasons is a prevailing lack of qualitative in-depth analysis of archival user needs [Cra03, Sin10] which would allow to analyze existing archival knowledge bases and to improve digital archival information systems [And04]. This requires, however, adequate, ontological and formal representations of the user needs towards archives.

The aim of the study[1] is to give empirical insight into the nature of user inquiries to archives and to investigate how patterns of inquiries can be reasonably represented in an ontological model in order to produce adequate answers. Such reasonable ontological representations of the research interest of the users as queries against an archival target world contribute to the creation of better documentation structures and better query facilities for archival information systems, for example, pattern-based [DKP00] query mechanism which would go beyond plain keyword searches.

In this paper, an overview of the research data and the methodology is given and the draft of one pattern, the *Documentation-Activity*, introduced. A brief example will demonstrate how existing EAD encoded archival data can be represented using this pattern.

2 Research Data

The term *reference question* refers to a request of a user to a staff member of a library or archive for information or assistance regarding the provision of any kind of information. Such a request can either be posed in person at a reference desk or remotely by

In: U. Kruschwitz, F. Hopfgartner and C. Gurrin (eds.): Proceedings of the MindTheGap'14 Workshop, Berlin, Germany, 4-March-2014, published at http://ceur-ws.org

[1] An extended version of this paper can be found in the preliminary proceedings of the CRMEX workshop (`http://www.ontotext.com/CRMEX`).

phone, mail, or e-mail. In this study, only written reference questions by mail or e-mail are being analyzed.

Archival reference questions capture an important phase of research: Expressing and formulating the wanted information or research interest as explicitly as possible by providing contextual information for another person. This kind of empirical research data contains a largely unfiltered information need of the user in his own words [DJ01] which constitutes a significant advantage over other methods of data collection like interviews or observation in existing information systems through, for example, log files, both of which elicit data biased by the interviewee or the preconditions of the information system.

Research data has been collected from the Federal Archives of Germany, the *Bundesarchiv*.[2] As a state archive, the Federal Archives are responsible for the permanent preservation and accessibility of federal archival documents such as files, papers, cartographic records, pictures, posters, films, sound recordings and machine-readable data.

User files hold physical copies and print-outs of letters or e-mails sent to the Bundesarchiv. The user files and the inquiries analyzed share a general historical and topical horizon which is Contemporary German History, understood as the history of the 19th and 20th century. Altogether, 236 user files have been selected. From these 236 initially selected user files 100 were available of which 60 contained at least one explicit or implicit information request as part of an inquiry by e-mail or letter. From these 60 user files, 546 single questions have been manually extracted based on the methodology outlined in the next section.

3 Methodological Approach

Archival reference questions have been largely neglected as research data. The study of Duff and Johnson [DJ01] is one of the few which looks at the type and structure of user reference questions. The study focuses on the types of questions and the types of elements used to contextualize the wanted information. Here, Duff and Johnson adapt a methodology for analyzing library reference questions based on the work by Grogan [Gro92] and Jahoda and Braunagel [JB80].

However, Duff and Johnson mainly focus on the *Aussageform* of the inquiries from an mostly archival point of view: First, they categorize the inquiries according to the *type of question*, for example, material-finding, fact-finding, or service request. Secondly, they systematize *given* and *wanted* information: The wanted information may be, for example, biographical information, location of a document, or general background information; the given information contextualizes the wanted information by, for example, proper names, place names, or a date.

The current study goes a step further and focuses on the *Erkenntnisform* of the inquiries, their epistemological form: The wanted information is interpreted regarding the *research interest* from a user point of view in order to describe reality in a way so that it fits the perceived epistemological interest of the user and his question. This ultimately means that the wanted information is determined more precisely by contextualizing it through explicit relations to the appropriate historical background as described by the given information. Through reasonable abstractions, the research interests is further generalized to common universals [MBG$^+$03, p. 8], i.e. generic relations and classes which have variations of themselves (e.g. *human being*) as opposed to particulars which have no variations of themselves (e.g. *Fritz*).

Regarding epistemological issues of the interpretation itself in relation to historical sciences or theory of history, the approach to interpretation taken here understands itself as *meta-theoretical*, similar to Gardin [Gar02] in the domain of archeology. The approach is agnostic to specific types of historical sciences but reflects patterns which can be considered applicable to general historical inquiry, for example, the pivotal role of actors and events and, in close relation to the archival target domain, the role of mostly written traces in the archives as evidence or source of information for historical investigations.

The CIDOC CRM [DOS07, DI08] is an ontological model which has been chosen as the means to formalize the results from the interpretations. One of the most important design principles of the CIDOC CRM is to represent the past as discrete events. Material and immaterial persistent items are present at events either as a concept or via a physical information carrier. History, therefore, is conceptualized as meetings of persistent items through events in space-times. Historical facts are described in terms of relations between universals. Since the model has been developed bottom-up from the analysis of a broad range of diverse cultural heritage ontologies, it has a strong empirical background and can be expected to be a suitable compromise between historical and archival conceptualizations.

This study adopts the methodology of the CIDOC CRM and tries if it either partially or completely covers this hypothetical ontology.

[2] A second, similar sample will be collected from the Norwegian National Archive.

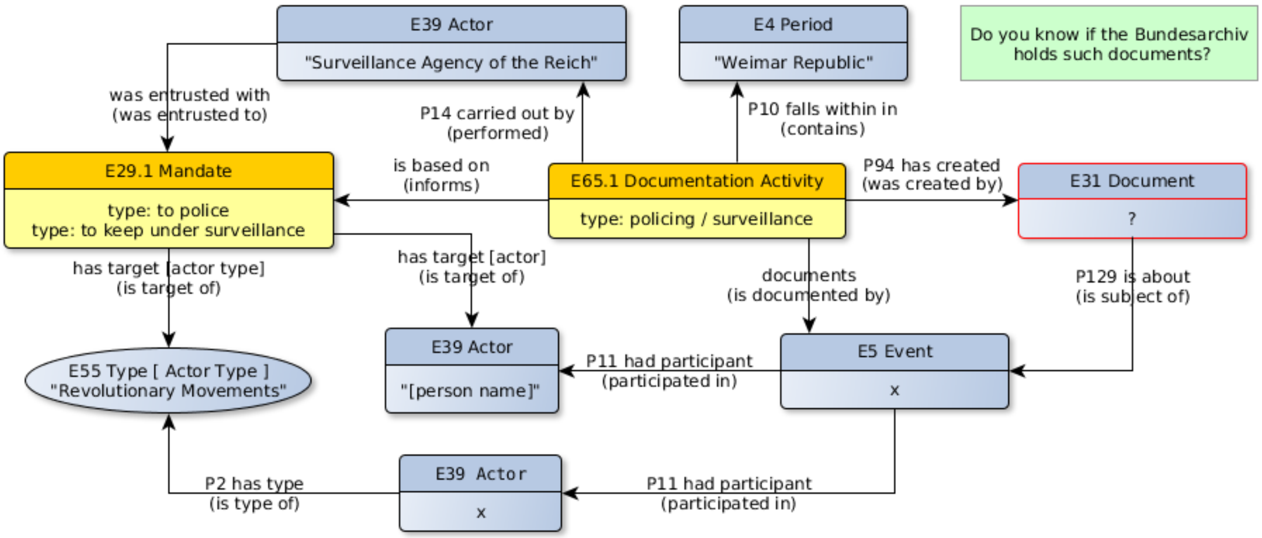

Figure 1: The example inquiry represented in the Documentation-Activity pattern.

4 Documentation Activity pattern

Preliminary results show that research interests found in inquiries can be reasonably represented as general patterns using CIDOC CRM. The *Documentation-Activity* pattern appears to be one of the most significant ones.

This pattern is the result of the interpretation of a broad range of inquiries and represents research interests targeted at documents which are the result of an activity[3] which documents events or, more specifically, observe the activities of people or groups: For example, the members of a parliamentarian committee document their meetings through minutes, or a secret agency observes the activities of a person through surveillance and generates a report.

The following question is a simple example for the interpretative analysis and formal representation of the research interest of an inquiry with CIDOC CRM.[4]

The context given in the inquiry is: *"One source I would like to consult are the police- and surveillance reports for the Weimar Republic which are about revolutionary movements. I would like to know what the surveillance agency of the Reich (or the ones of the Länder) had to say about [person name]."*[5]

The question asked in the inquiry reads: *"Do you know if the Bundesarchiv holds such documents?"*

The first interpretation step asks if there are probable and adequate answers to the question with regard to the domain of historical inquiry but also to the archival domain. Here, the user is looking for reports which are the result of a policing or surveillance activity targeted at a specific type of group ("revolutionary movements") or at a specific person ("[person name]"). In that way, this question could be even seen as a two-fold question. The result of these policing or surveillance activities are documents about the activities of the aforementioned actors. Such documents are routinely products of a governmental institution and are now stored in an archive. The user wants to know if such documents are available in the Bundesarchiv. Therefore, the information the user wants are pointers to appropriate documents, for example, call numbers of files likely to contain relevant documents.

The second interpretation step comprises the translation of the question, its context and its interpretation to the CIDOC CRM. The two-fold question can be represented as shown in figure 1. This is a simplified representation expressing the formal basic structure of an answer adequate to satisfy the wanted information or the research interest.[6] The interpretation of the question is evident and materialized by the documentation activity[7] in the center of the figure. The documentation activity is seen as being implicit in the historical reality referred to in the question: The police- and surveillance reports have been created during an event, or a series of events, which "documented" some other events and which are qualified by the participation of

[3] In CIDOC CRM, E7 Activities are sub-classes of E5 Events.

[4] Note, that the inquiry has been translated from German to English by the author of this paper.

[5] The name of the person referred to has been rendered anonymous.

[6] The implicit question for pointers to documents, for example, a set of call numbers, is not the point when translating to CIDOC CRM but the *context* of the documents of interest. Identification for retrieving the actual physical document is not in the scope of this ontological model.

[7] An extension to the CIDOC CRM currently deemed necessary.

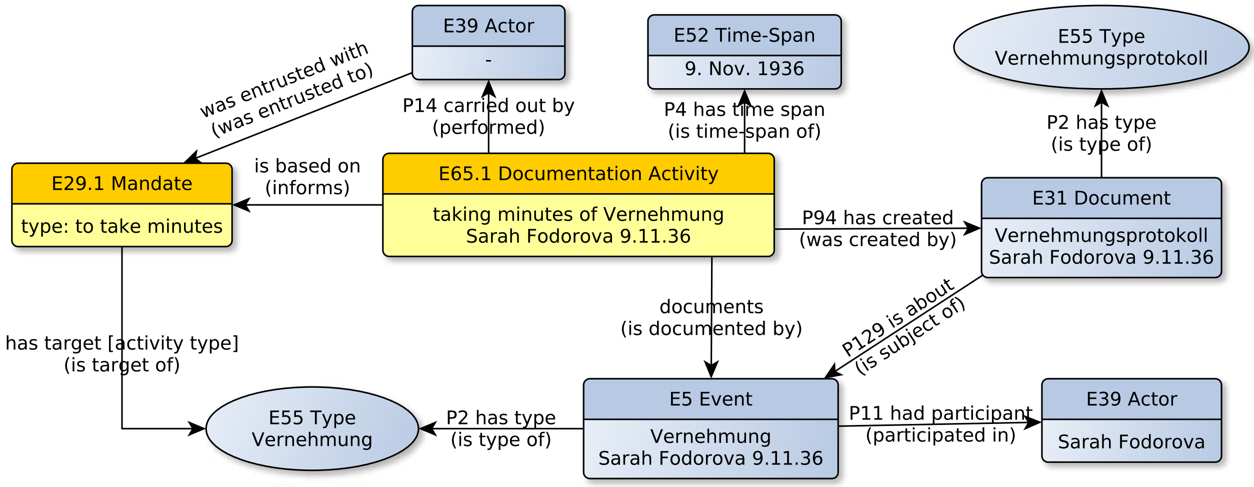

Figure 2: The information from the <unittitle> represented explicitly.

an actor ("[person name]") or a specific type of group ("revolutionary movements"). The documentation activity is following a mandate which captures a specific type of "documented plans (...) for deliberate human activities [CDG+11, p. 15]."

Most importantly, mandates[8] specify or govern documentation activities. In the case of the two-fold question the mandate has a specific type of group as its principle target and at the same time aims at a specific actor. Furthermore, the mandate is assigned to an actor, in this case an institution, who carries out the actual documentation activity which, as the last relevant contextual information, falls within the historical period of the Weimar Republic. Documents which are the result of this constellation are relevant documents and may adequately answer the user's two-fold question.

This brief example demonstrates how the research interest of inquiries can be formally represented in an abstract ontological model. The next section will show how such a pattern could be instantiated with empirical data from a digital archival aid.

5 AKM and EAD

The *Archival Knowledge Model* (AKM) comprises a set of such patterns like the Documentation-Activity. As a Conceptual Reference Model it can be used to analyze and to query archival knowledge bases. Tzompanaki and Doerr [TD12] show how large and complex semantic networks may be queried using CIDOC CRM. Especially in cases where relevant documents can be expected to be distribute among records or holdings, such patterns would provide relevant access points and contexts to retrieve documents.

Here, a brief example shall demonstrate how archival finding aids encoded with EAD could be analyzed whether they provide sufficient implicit or explicit information to adequately answer typical user queries.

The *Encoded Archival Description*[9] (EAD) standard is the de facto standard for the digital encoding of archival aids. One of the essential information entities in a finding aid encoded in EAD is the element `<unittitle>` which typically holds the "name of the described materials"[10] at any level of the descriptive tree.

The following XML snippet is taken from the existing EAD finding aid *Roter Koffer*[11] from the Bundesarchiv. In this case it represents a quite informative but yet typical entry in an archival finding aid giving the title of a file: `<unittitle>Vernehmungsprotokoll Sarah Fodorova vom 9. Nov. 1936</unittitle>`.

This `<unittitle>` contains a lot of implicit information: There has been an interrogation (*Vernehmung*) of a person named Sarah Fodorova on the 9.11.1936 which has been documented by minutes (*Vernehmungsprotokoll*) which are now stored in the file.

Figure 2 shows an exemplary instantiation (of parts) of the Documentation-Activity pattern with the

[8] This class is another proposed extension to the CIDOC CRM.

[9] `http://www.loc.gov/ead/`

[10] `http://www.loc.gov/ead/tglib/elements/unittitle.html`

[11] "Roter Koffer" translates to "Red Suitcase". For background information on this holding confer: `http://www.bstu.bund.de/DE/Wissen/Aktenfunde/Roter-Koffer/roter-koffer_inhalt.html`

information from the <unittitle>. In this representation the information is explicit and formalized according to a pattern which is relevant to a broad range of information needs of typical user inquiries.

The example also shows that even though the AKM may seem complex, sufficient semantics can be expected to exist in literal information values. The patterns documented in the AKM are evidently implementable by data structures improved accordingly.

Lastly, the intellectual work for the archivist when creating the title remains the same when he serves the seemingly more complex pattern.[12] On the contrary, his intellectual work is preserved in a relevant and explicit representation while it would be lost in a plain literal text.

6 Conclusion

In terms of its research data and methodological approach the research introduced in this paper appears to be rare among studies of the information behavior of archival users. The study and its research data are empirical in nature, however, the employed methodology has a strong interpretative approach. Archival reference questions are a research data which is difficult to obtain and analyze, however, the interpretative analysis and formalization of written natural language questions from users to archives, as has been tried to demonstrate, constitute a valuable source for obtaining meaningful data on original user needs. Only if we gain a significant and deeper understanding and consensus on archival user needs in general we will be able to build a new generation of more sophisticated pattern-oriented (archival) information systems for the (archival) users.

References

[And04] Ian G. Anderson. Are you being served? historians and the search for primary sources. *Archivaria*, (58), 2004.

[CDG+11] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead, Matthew Stiff, and ICOM/CIDOC CRM Special Interest Group. Definition of the CIDOC conceptual reference model (version 5.0.4): Produced by the ICOM/CIDOC documentation standards group, continued by the CIDOC CRM special interest group, 2011.

[Cox08] Richard Cox. Revisiting the archival finding aid. *Journal of Archival Organization*, 5(4), 2008.

[Cra03] Barbara Craig. Perimeters with fences? or thresholds with doors? two views of a border. *American Archivist*, 66(1), 2003.

[DI08] Martin Doerr and Dolores Iorizzo. The dream of a global knowledge network: A new approach. *Journal on Computing and Cultural Heritage*, 1(1), 2008.

[DJ01] Wendy M. Duff and Catherine A. Johnson. A virtual expression of need: An analysis of e-mail reference questions. *American Archivist*, 64(1):43–60, 2001.

[DKP00] Garett O. Dworman, Steven O. Kimbrough, and Chuck Patch. On pattern-directed search of archives and collections. *Journal of the American Society for Information Science*, 51(1), 2000.

[DOS07] Martin Doerr, Christian-Emil Ore, and Stephen Stead. The CIDOC conceptual reference model: A new standard for knowledge sharing. ER2007 tutorial. *Challenges in Conceptual Modelling: Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling, ER 2007, Auckland, New Zealand, November 5-9, 2007*, 83, 2007.

[Gar02] Jean-Claude Gardin. Archaeological discourse, conceptual modelling and digitalisation: An interim report of the logicist program. *The Digital Heritage of Archaeology: Computer Applications and Quantitative Methods in Archaeology, Proceedings of the 30th Conference, Heraklion, Crete, April 2002, CAA 2002*, 2002.

[Gro92] Denis Grogan. *Practical Reference Work*. Library Association Publishing, London, 2. edition, 1992.

[JB80] Gerald Jahoda and Judith Schiek Braunagel. *The Librarian and Reference Queries: A Systematic Approach*. Library and information science. Academic Press, New York, 1980.

[MBG+03] Claudio Masolo, Stefano Borgo, Nicola Guarino, Alessandro Oltramari, and Luc Schneider. WonderWeb deliverable d17. the WonderWeb library of foundational

[12] The "mechanical" effort might differ in that it is quick and easy to simply type in a literal text. However, this is a question of implementation and of proper tool design for the creation of archival aids.

ontologies. preliminary report. Deliverable D17, May 2003.

[MH01] Angelika Menne-Haritz. Access: The reformulation of an archival paradigm. *Archival Science*, 1, 2001.

[Sin10] Donghee Sinn. Room for archives? use of archival materials in no gun ri research. *Archival Science*, 10(2), 2010.

[TD12] Katerina Tzompanaki and Martin Doerr. A new framework for querying semantic networks. San Diego, 2012.