Representation of Archival User Needs using CIDOC CRM

Steffen Hennicke[1]

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

Abstract. This paper stems from an ongoing dissertation project and demonstrates how the CIDOC CRM is used 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 query archival or historical knowledge bases in order to provide more adequate answers and to enable more relevant discovery facilities. For this purpose, 330 reference questions to the German Federal Archive are being analyzed and patterns found translated to the CIDOC CRM and appropriate extensions. In particular, the paper introduces the methodological approach to the interpretation of user questions and the draft of a prominent pattern called *Documentation-Activity*.

Keywords: CIDOC CRM, archival reference questions, access to archives, archival user needs, Archival Knowledge Model

1 Introduction

The main means for discovering [1] and accessing primary sources in an archive are finding aids and holding guides supported by the expertise of archivists. These archival aids are descriptive tools which are meant to help the user to locate and discover relevant materials in the enormous and ever growing amounts of rich *information potentials* [2] in archives. The conceptualization of these descriptive tools as well as respective digital encoding standards like the *Encoded Archival Description*[1] (EAD) are based on elaborated and historically grown archival principles and models but their design is less informed by explicit knowledge about the information needs of archival users [3] due to a prevailing lack of qualitative in-depth analysis of archival user needs [4][5]. Such studies are, however, a crucial cornerstone for the improvement of existing and future digital archival information systems [6].

The hypothesis on which this paper rests is that it is not necessary, and even not desirable, to change the archival description itself and related metadata schemas, but, instead, it is possible, in principle, to supplement existing archival

[1] http://www.loc.gov/ead/

and historical knowledge bases with an ontological model which matches typical patterns from user inquiries to archives. Such an ontological model would make knowledge explicit and add relevant context which is necessary to adequately answer typical user questions and to create better discovery systems. Furthermore, such ontological models enable empirically qualified assessments of metadata schemas for archival information systems but also of archival cataloging rules.

The general research question, therefore, is if there is a hypothetical ontology which can represent user inquiries and their probable interpretations as formal queries against a model of the archival target world that would adequately answer the inquiry or its implicit purpose. The result of this analysis is an ontological model which represents inquiry patterns of different abstraction levels to archives in the form of queries to this ontology. The CIDOC CRM[2] has been chosen as the ontological target model mainly for its strong empirical foundation and event-based conceptualization of historical processes. Written *reference questions*[3] from the German Federal Archives, the *Bundesarchiv*, have been chosen as research data. This type of research data has been largely neglected in the analysis of user needs in the archival domain, although they document a mostly unfiltered information need in the users own words.

A brief literature review will establish the general research context followed by a short introduction of the research data and the methodological approach to the analysis of questions.[4] The focus of this paper lies on the demonstration of the interpretative translation of natural language questions to a common ontological representation. Two examples will demonstrate how shared patterns in user questions and their probable interpretations can be translated to an ontological model, the *Archival Knowledge Model* (AKM), covering and extending the CIDOC CRM. The specific pattern presented in this paper is called *Documentation-Activity* which proposes two new classes as extensions to the CIDOC CRM. It is important to bear in mind that all results presented in this paper are preliminary and research is ongoing.

2 Research Context

The limitation on "simple answers to clear cut, search term-based questions" [7] is one of the core problems of today's information systems on the Web. Pattern-oriented retrieval could describe many more complex questions whose answers go beyond the capacity of simple querying [8]. This limitation poses a significant

[2] http://cidoc-crm.org/

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

[4] For more details on the dissertation, please confer the extended abstract which will be presented at the Doctoral Consortium of the TPDL 2013 and published in the conference proceedings.

barrier to more sophisticated and integrated information systems. Part of this problem is a prevalent focus on traditional library cataloging and methodology in describing and contextualizing objects of interest. At the same time, today "the key challenge of organizing information is to construct systems that aid understanding, contextualizing, and orienting oneself within a mass of resources" where models help to bridge a semantic gap between the formalizations in information systems and the conceptualizations of scholars [9]. Instead of a *Global Knowledge Network* [7], mostly isolated "silos of information" exist which all employ their own idiosyncratic structures and data encodings. The Semantic Web addresses these issues in its research agenda. However, this agenda suffers from an "almost exclusive focus on 'terminology' rather than 'ontological structures'" resulting in the neglect of fundamental and complementary lines of research [7]. One such missing line of research is how typical user questions are formally structured. The systematic and in-depth analysis of original user questions from different stages of the research process is important and has the potential to provide, for example, necessary information on query mechanisms or adequate granularity of ontologies [7].

Discovery is one of the most important and re-occurring stages in research processes especially distinctive for historical inquiry in archival settings. As already mentioned, research in the archival domain exhibits a lack of in-depth user studies [10]. The study of Duff and Johnson [11] is one of the few which looked at the type and structure of user reference questions to archives.[5] Regarding the domain of historical research, Case [15] concluded that history "may be less well served by classification and indexing than any other academic field" and that the "accomplished scholar - and particularly the historian - is not often aided by the disciplinary boundaries that library classification schemes enforce." Instead of the "disciplinary model of a body of knowledge, subdivided by place and period", the so called "problem-oriented model" should be used as the basis for the design of future tools and services for historians. At the same time, Case correctly points out that it is not viable to fundamentally change documentation practices and reorder collections of archives and libraries but that special services and tools might be able to bridge (semantic) gaps between the user and existing knowledge bases.

3 Research Data

The *Bundesarchiv* is the Federal Archives of Germany who are responsible for the permanent preservation and accessibility of federal archival documents from the civil and military archives of the Federal Republic of Germany (since 1949) and its predecessors. In addition, significant documents of private origins and from political parties, associations and societies are kept in the archive. The number of written inquiries to the Bundesarchiv amounts to roughly 60,000 per

[5] Similar studies are, for example, from Collins [12], Conway [13], and Gagnon-Arguin [14].

year, based on the numbers from 2008 and 2009.[6] The Bundesarchiv has granted supervised access to their user files which contain a physical documentation of the correspondence and interaction between a user and the archive. Each user file carries an identification number which is retrievable through a database system offering a small range of search facets[7] related to the user and associated user files. Based on these facets a sample of 196 user files was retrieved, which was further complemented by a special selection of 40 rich user files which had been collected by the head of the department *Stiftung Archiv der Parteien und Massenorganisationen der DDR im Bundesarchiv* (StA). The sample of user files shares as a general historical and topical horizon Contemporary German History (19^{th} and 20^{th} century) and contains rich and challenging inquiries. The sampling process was informed by educated assumptions, professional advice of archivists, and skimming through user files. The collection was stopped when the questions extracted from the user files appeared to exhibit no new qualities or substantial variances. Reliable information about the users' background was not available.

In total, 236 user files have been selected from which 100 were available for further study. Only 60 user files contained at least one explicit or implicit information request as part of an inquiry by email or letter. From these 60 user files, 546 single questions were extracted and pre-analyzed[8] according to the methodology of Duff and Johnson [11] with very similar results. Regarding the type of question, 260 questions were of type "explicit" or "implicit resource discovery" (material-finding, specific form, specific item, consultation), 70 questions were "factual", and 216 questions consisted of "administrative/directional", "user education", or "service request" questions. The questions of the type "resource discovery" and "factual", in total 330 questions, are part of the discovery stage in the research process and are currently being analyzed as described in the following sections.

4 Methodological Approach

The methodological approach taken in this study goes beyond the analysis of the mere utterance level and syntactic structure of the inquiry and focuses on the interpretation of the questions. Here, the sense of an inquiry is interpreted in order to discover the implicit questions with regard to a certain domain of discourse. In the scope of this work, two domains of discourse are being distinguished: the *archival domain* of record keeping and the *domain of historical*

[6] http://www.bundesarchiv.de/oeffentlichkeitsarbeit/publikationen/taetigkeitsberichte/

[7] This includes, for example, the general purpose of the inquiry as given by the user on the user management form, a general subject and time frame of the inquiry's topic, or the department initially responsible for processing the inquiry which allows concluding on the origins of the archival material. However, it is important to note that these classifications are coarse and not meant for precise retrieval of user files based on these search facets.

[8] The publication of the results is in preparation.

inquiry for which traces and evidence can be expected to be found in the archive. These two domains constitute the epistemological baseline for the interpretation of the inquiries: What might the user need to know in order to satisfy his research interest? Reality is then described in a way so that it fits the perceived epistemological interest of the user and his question. This process is necessarily an act of interpretation and relies on educated intuition regarding both domains and necessarily filters probable implicit questions. It does not, to be sure, aim at "truthful" models in terms of some perceived "objective" meaning or structure of a question. Regarding such epistemological issues of interpretation in relation to historical science and theory of history, the approach to interpretation taken here understands itself as *meta-theoretical*, similar to Gardin [16] in the domain of archeology. It is agnostic to specific types of historical sciences but reflects patterns applicable to general historical inquiry.

The patterns which are identified in the questions are modeled in CIDOC CRM which describes historical facts in terms of possible relations between universals. It is the result of an empirical analysis of existing conceptualizations of the cultural-historical world in the form of metadata structures. 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 [17]. The historical-archival domain of the analyzed inquiries is in the scope of the CIDOC CRM. For these reasons, its methodology is adopted in this work and it is tried to identify whether the CIDOC CRM will completely or partially cover the hypothetical ontology. In the latter case, appropriate extensions to the CIDOC CRM will be proposed.

Formally, an ontology engineering approach is employed in that the inquiries and their interpretations are being translated to an ontological model based on the CIDOC CRM and appropriate extensions.

5 Translating Patterns of Questions to CIDOC CRM

Two examples will motivate how questions are being analyzed and how their interpretation is formally represented in an ontological model based on the CIDOC CRM. An inquiry typically consist of contextual information and one or more direct or indirect questions.[9]

5.1 Example 1

Context: *One source I would like to consult are the police- and surveillance reports for the Weimar Republic which are about revolutionary movements. I would*

[9] All questions have been translated from German to English by the author of the paper. Text in square brackets has been added either to make named entities anonymous or to clarify the meaning of certain paragraphs. Finally, red borderlines indicate the entity at which a question is targeted.

like to know what the surveillance agency of the Reich (or the ones of the Länder) had to say about [person name].

Question 1: *Do you know if the Bundesarchiv holds such documents?*

Question 2: *Which agency of the Reich was responsible for the surveillance of the revolutionary movements? The Reich or the Länder?*

The *given* elements in these two questions and their context are the name of a specific actor ("[person name]"), the type or function of a group ("revolutionary movements"), the type or function of a legal body ("surveillance agency of the Reich"), the type or function of documents ("police- and surveillance reports"), and the name of a period ("Weimar Republic").

The interpretation of the questions can be structured into two principle steps. The first one is concerned with the *wanted* information asking for the research interest of the user's question: Which are probable or adequate answers to the question with regard to the domain of historical inquiry but also to the archival domain?

In the case of the first question 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, the first question could be even seen as a two-fold question. The results of these policing or surveillance activities are documents about the activities of the aforementioned actors. Such documents as routinely products of a governmental institution 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 question in the example is a fact-finding question. It operates with the same given information but asks for a different wanted information. The user wants to know which agency was generally responsible for surveillance activities targeted at a specific type of actor. He is inquiring for a name of one or more legal bodies. The word "responsible" is important because it stresses the fact that whatever agency conducted the surveillance activities did so following a mandate which formally delegated said responsibility to the agency.

The second interpretation step comprises the translation of the question, its context and its interpretation to the CIDOC CRM. The CIDOC CRM suggests that historical facts and entities are related to each other through events which form the world lines in history. Therefore, the second interpretation step asks how the given and wanted information entities relate to each other.

The first two-fold question can be represented in CIDOC CRM 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.[10] The interpretation of the question is evident and materialized by the

[10] 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.

documentation activity in the center of the figure. This class is a proposed extension to the CIDOC CRM and will be introduced in more detail later on. 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 events which are qualified by the participation of 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".

Most importantly, mandates specify or govern documentation activities. This class is another proposed extension to the CIDOC CRM and will also be introduced in more detail later on. In the case of the first 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 in 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 first two-fold question.

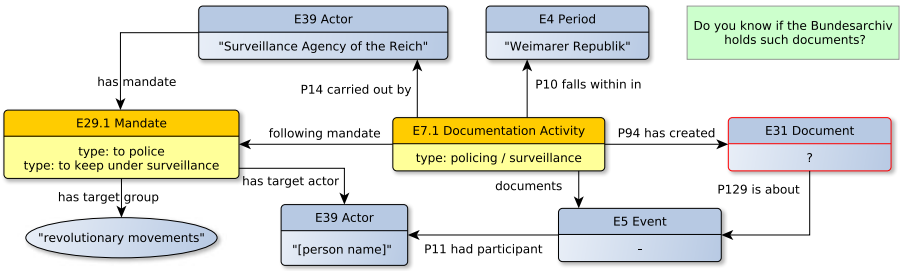

Fig. 1. Question 1 in AKM

Figure 2 shows the translation of the interpretation of the second question to the CIDOC CRM. An adequate answer can be modeled within the same pattern as for the first question. In this case the wanted information is the name of an actor who had the mandate to police or to keep under surveillance revolutionary movements during the Weimar Republic.

These two questions and their representations in CIDOC CRM show a common core pattern which is grouped around a documentation activity which documents events and which is following a specific mandate. This relation between documentation activity and mandate is essential. It can be identified in many other questions through interpretation.

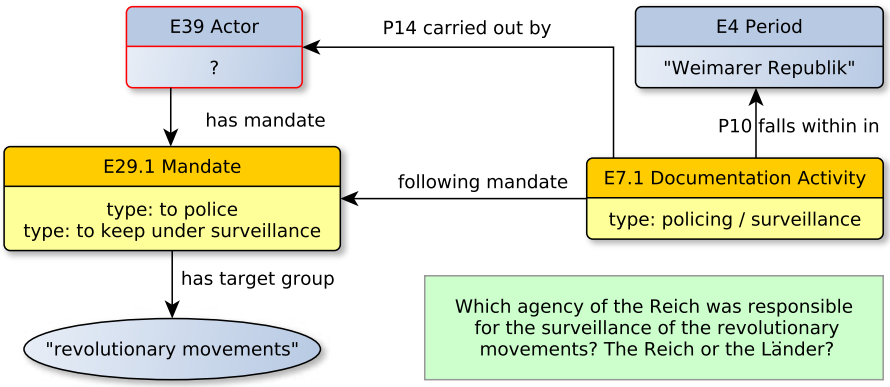

Fig. 2. Question 1 in AKM

5.2 Example 2

The second example shows that seemingly different questions exhibit very similar patterns and that documentation activities based on mandates may cover a broad range of different types of activities. Furthermore, some finer notions like self-documentation and documentation of others are introduced in this example.

Context: *In 1980, a delegation of the FDGB lead by Harry Tisch laid down a wreath of flowers in Oradour. The visit was part of a trip of the FDGB to France (demonstration in Limoges, reception and meeting with the FKP and CGT in Paris). At this time, Tisch was also a member of the Politbüro of the ZK of the SED.*

Question 1: *Where can documents be found about the planning [of this trip]...*

Question 2: *...and the report on this trip?*

Question 3: *In your opinion, has such a trip been discussed or, at least, been approved in the ZK?*

The first question asks for documents about the planning of the trip to France while the second question asks for the report on this trip. In both cases the documents refer to the same event "Trip to France" but they are the result of two distinct activities. The first one, the planning activity, happens prior to the actual trip and does not directly document the trip but series of planning events. The second documentation activity, the reporting activity, produces one or more documents which report on the trip event itself. Both questions ask for pointers to documents as the result of their respective documentation activity.

Figure 3 combines the first and second question and their interpretations. The documents are the result of documentation activities which document events which were involved in the planning of the trip to France. In the case of the second question, the documents are the result of a documentation activity reporting on the event "Trip to France". Necessarily, both documentation activities followed a mandate to do so and were carried about some actor.

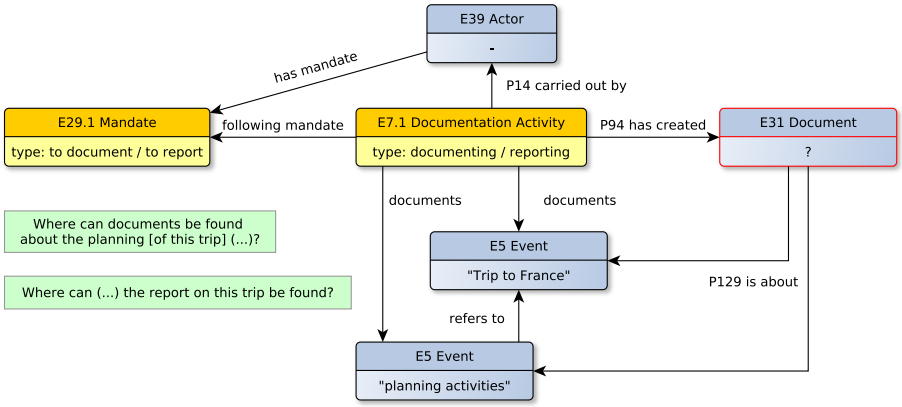

Fig. 3. Questions 1 and 2 in AKM

The third question should adumbrate some more difficult issue in terms of interpretation and translation of questions. The question asks if a specific actor, the "ZK"[11], had discussed or approved a specific event, the trip to France.

First of all, it is important to remember that the patterns are about the general and generic relations between certain entities and not about the many specific qualities of these connections: it is not relevant if the relation between a document and an event is one of "discusses" or "approves" but that, on the most generic semantic level, it is a relation of "aboutness". It is the genuine task of the researcher to read and interpret the documents in order to find out about the qualitative aspect if the ZK did in fact "discuss" and, even more, did "approve" something. The pattern is a means for the researcher to discover potentially relevant documents. One tentative inference which might be drawn from a knowledge base which instantiates this ontological model is that the ZK, or more precisely some members of this group, must have had knowledge of the event "Trip to France".

Therefore, relevant documents for an adequate answer include those ones which are about events during which the actor ZK was present and which in some way refer to the event "Trip to France". An example for such an event could be the planning event from the first question. Figure 4 shows another possible scenario where the ZK carried out a committee meeting during which the trip to France has been mentioned and which has been documented through minutes.

Again, the minutes are the result of a documentation activity which follows a mandate to take minutes. In this case, the ZK is the actor who not only follows this mandate and carries out the documentation but also conducted the event which is being documented. This is a kind of *self-documentation* as opposed

[11] "ZK" is the abbreviation for *Zentralkomittee* ("central committee") which belonged to the *Sozialistische Einheitspartei Deutschland* (SED).

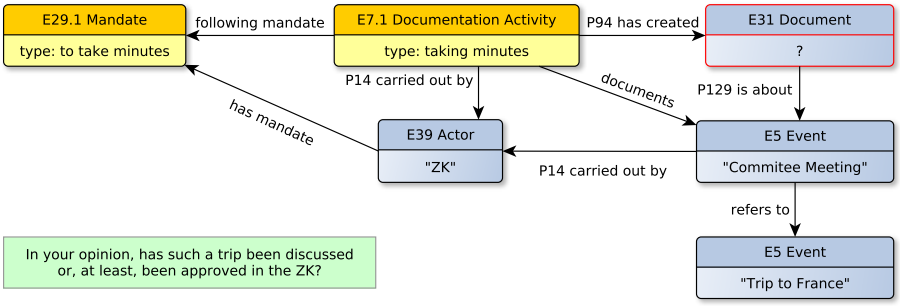

Fig. 4. Questions 3 in AKM

to the documentation of others in the case of surveillance and will be briefly discussed in the next section.

6 The Documentation-Activity Pattern

The examples previously discussed exhibit a shared pattern which is able to accommodate a broad range of different questions and their probable interpretation in terms of adequate answers. This section will introduce the current draft of the *Documentation-Activity* pattern as shown in figure 5. So far, this pattern appears to be one of the most prominent and complex results from the analysis of the user questions.[12]

At the core of this pattern resides a new proposed class *E7.1 Documentation Activity*. This new event class is an extension to the CIDOC CRM in order to appropriately capture the essentials of activities which, literally speaking, document *E5 Events* and which create one or more *E31 Documents*. It is a sub-class of *E65 Creation* and not of *E7 Activity* because a characteristic feature of the documentation activity is the creation of documents and only events of the type *E65 Creation* "result in the creation of conceptual items or immaterial products" through *P94 has created*. Furthermore, the scope of *E7.1 Documentation Activity* is more specific than that of *E65 Creation* in that documentation activities document *E5 Events* and, most importantly, follow a mandate. The representation of the fact that a documentation activity follows a mandate led to the introduction of a new property called *follows mandate*.

The *E29.1 Mandate* is the second proposed extension to the CIDOC CRM as a sub-class of *E29 Design or Procedure*. The mandate formulates the principle scope of application for documentation activities in that it specifies who has the mandate to execute the documentation activity and which specific actors, types

[12] While the analysis of the questions is on-going and no reliable evidence based on the current research sample can be provided at this point, an estimate of at least 30% of all questions in the sample might be adequately covered by this pattern either partially or in full.

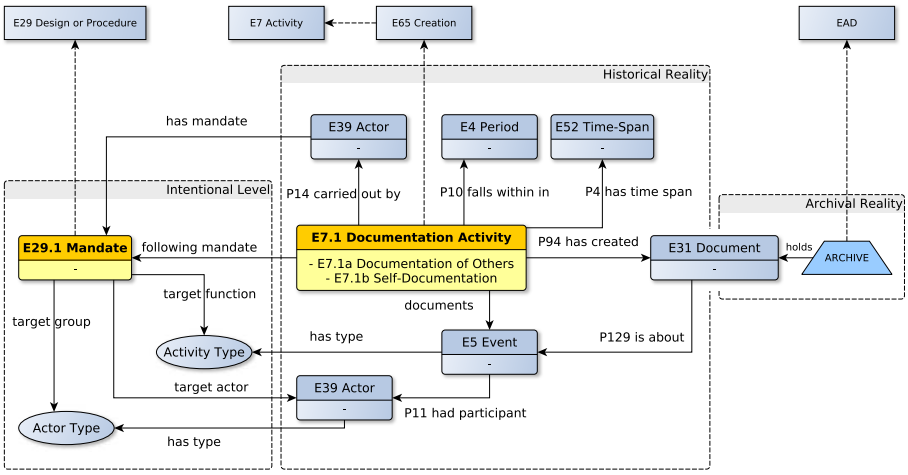

Fig. 5. The current draft of the Documentation-Activity pattern

of actors, or types of activities may be the target. In order to appropriately describe these target relations new properties – *target function*, *target group*, *target actor*, and *has mandate* – have been introduced.

The *E7.1 Documentation Activity* and the *E29.1 Mandate* related through *follows mandate* constitute the essential core of the identified common pattern: The documentation of events according to standing mandates producing documents which can be found in the archive. This mandate-based documentation (*auftragsgemäße Dokumentation*) can not be adequately represented with *E65 Creation* and *E29 Design or Procedure*. The pattern allows to draw conclusions on the probability that specific types of events have been documented and that traces can be expected in the archive.

The documentation activity and its contextual classes can be seen as being part of a description of the *historical reality* as given in the user's question. The mandate, on the other hand, belongs to an *intentional level* (*Absichtsebene*) where principles are formulated which are meant to formally govern the historical reality and which might find their expression in documents. These documents are the point where this ontological representation of the historical reality would intersect with the one of the *archival domain* of record keeping as indicated in figure 5. It is important to note that an *E31 Document* is not a physical item but "comprises identifiable immaterial items that make propositions about reality". A physical materialization of an *E31 Document* in the archive may be an *E33 Linguistic Object* which "comprises identifiable expressions in natural language or languages". Here, a model of expressions of documents in the archive is not included.[13]

[13] Cf. [18] for an approach to mapping EAD to the CIDOC CRM.

The analysis of the questions is on-going and changes to the pattern might occur and there are other aspects which appear to be relevant. The *official* and *unofficial* nature of a document, for example, seems to be another important aspect. This point cannot be discussed in any detail in this paper, however, if a document is official or unofficial is most likely determined by the circumstances of its publication. As already mentioned, the examples also show cases in which the documentation activity is carried out by the same actor who is also responsible for the documented activity. This is a kind of self-documentation giving an "official account" (*Rechenschaftsbericht*) such as proceedings, government statements etc. In the Documentation Activity pattern this can be expressed by two principle sub-types *E7.1a Self-Documentation* and *E7.1b Documentation of Others*.

7 Conclusion

This paper introduced the draft of the *Documentation-Activity* pattern which is part of the *Archival Knowledge Model* (AKM). The AKM is an ontological model which comprises representations of general patterns found in archival user inquiries and their interpretations.

Such an ontological model can help to bridge the semantic gap between traditional archival documentation and organizing principles and the conceptualizations employed by different kinds of users and support building search and discovery systems which are able to better respond to pattern-oriented questions. As a formal model, the AKM could also inform the design of archival metadata schemas or new archival "cataloging rules" as, for example, that titles of series or files should not be plain text but structured according to patterns like the Documentation-Activity pattern.

References

1. Unsworth, J.: Scholarly primitives: What methods do humanities researchers have in common, and how might our tools reflect this? (2000)
2. Menne-Haritz, A.: Access: The reformulation of an archival paradigm. Archival Science **1** (2001)
3. Cox, R.: Revisiting the archival finding aid. Journal of Archival Organization **5**(4) (2008)
4. Craig, B.: Perimeters withfences?or thresholds with doors? two views of a border. American Archivist **66**(1) (2003)
5. Sinn, D.: Room for archives? use of archival materials in no gun ri research. Archival Science **10**(2) (2010)
6. Anderson, I.G.: Are you being served? historians and the search for primary sources. Archivaria (58) (2004)
7. Doerr, M., Iorizzo, D.: The dream of a global knowledge network: A new approach. Journal on Computing and Cultural Heritage **1**(1) (2008)
8. Dworman, G.O., Kimbrough, S.O., Patch, C.: On pattern-directed search of archives and collections. Journal of the American Society for Information Science **51**(1) (2000)

9. Shaw, R.: Information organization and the philosophy of history. Journal of the American Society for Information Science and Technology **64**(6) (2013)
10. Harris, C.: Archives users in the digital era: A review of current research trends. Dalhousie Journal of Interdisciplinary Management **1** (2005)
11. Duff, W.M., Johnson, C.A.: A virtual expression of need: An analysis of e-mail reference questions. American Archivist **64**(1) (2001) 43–60
12. Collins, K.: Providing subject access to images: A study of user queries. American Archivist **61**(1) (1998)
13. Conway, P.: Partners in Research: Improving Access to the Nation's Archive. Archives & Museum Informatics, Pittsburgh (1994)
14. Gagnon-Arguin, L.: Les questions de recherche comme matriau dtudes des usagers en vue du traitement des archives. Archivaria **46**(1) (1998)
15. Case, D.O.: The collection and use of information by some american historians: A study of motives and methods. The Library Quarterly **61**(1) (1991)
16. Gardin, J.C.: 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)
17. Doerr, M.: The CIDOC conceptual reference module: An ontological approach to semantic interoperability of metadata. AI Magazine **24**(3) (2003)
18. Bountouri, L., Gergatsoulis, M.: Mapping encoded archival description to CIDOC CRM. Proceedings of the First Workshop on Digital Information Management (2011)