DELIVERABLE

Project Acronym: DM2E

Grant Agreement number: ICT-PSP-297274

Project Title: Digitised Manuscripts to Europeana

D3.4 – Research Report on DH Scholarly Primitives

Revision: Final 2.0

Authors:

Steffen Hennicke (UBER)
Stefan Gradmann (KUL)
Kristin Dill (ONB)
Gerold Tschumpel (UBER)
Klaus Thoden (MPIWG)
Christian Morbindoni (Net7)
Alois Pichler (UiB)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X

Revision history and statement of originality

Revision	Date	Author	Organisation	Description
0.1	12.1.15	Steffen Hennicke, Kristin Dill, Gerold Tschumpel Klaus Thoden	UBER, ONB, MPIWG	Initial draft
0.2	19.1.15	All	UBER, ONB, MPIWG	Revision
0.3	22.1.15	All	UBER, ONB, MPIWG	Revision
0.4	26.1.15	All	UBER, ONB, MPIWG	Revision
0.6	31.1.15	Steffen Hennicke	UBER	Minor changes
0.7	05.2.15	Violeta Trkulja, Steffen Hennicke	UBER	Minor corrections
0.8	09.2.15	Violeta Trkulja, Vivien Petras	UBER	Final revision
0.9	09.2.15	Steffen Hennicke	UBER	Final revision
1.0	10.2.15	Violeta Trkulja	UBER	Approval of Final 1.0
1.1	24.03.15	Steffen Hennicke	UBER	Changes in chapter 2.2 and 2.6
2.0	02.04.15	Violeta Trkulja	UBER	Approval of Final 2.0

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

List of Tables

List of Figures

1 Introduction

This is the last deliverable of Digitised Manuscripts to Europeana (DM2E) Work Package 3 (WP3) – "Research Report on DH Scholarly Primitives". It presents the results of Task 3.4 the "Background research on Scholarly Primitives". The Research Report (D3.4) addresses the following principle research questions established in the DoW: "What are the "functional primitives" of the digital humanities?", "What kinds of "reasoning" do digital humanists want to see enabled by the data and information available in Europeana together with those that are currently not part of the Europeana portal?" and "Which types of operations do digital humanists expect to apply to Europeana data and do they expect these to be offered by Europeana (API based) or from (external) third parties?"

We decided to answer these questions starting from the point of view of the framework of our project. This means that we used the Linked Data approach, the tools developed in WP3, i.e. Pundit, Ask, and the content from WP1. The overall guiding research question has therefore been to explore "how Linked Data based digital tools and data can support, facilitate, or enhance humanist work practices?"

Since this is a wide and complex topic and the scope of the task is broad, we approached it from several angles. This includes top-down conceptual research, for example regarding the Scholarly Domain Model (SDM), and bottom-up empirical research consisting of interviews and experiments with scholars. Our various methods for data collection included desk research, discussions with the members of the Digital Humanities Advisory Board[1] (DHAB) and in the WP3 working group, interviews, experiments and workshops, surveys and questionnaires.

Work on the task began with desk research and discussions within the project and with the DHAB. Based on these initial research, we devised the first version of the Scholarly Domain Model during the first year. This first version has been presented and discussed on numerous occasions (cf. 8. Appendix: Related Publications and Presentations) and then subsequently further revised. In particular, in order to collect empirical evidence regarding the adequacy of our approach, we conducted semi-structured interviews with humanists (cf. "Report on Interviews"). The data collected during these interviews and the continuous research into the matter resulted in the final version of the SDM on which we report in the section "Scholarly Domain Model".

Part of the continuous research was a series of experiments conducted with Pundit and its components during the last year of the project. In the beginning of 2014, we began preparing experiments with Pundit, particularly in order to investigate the Scholarly Activity Annotating, i.e. how exactly do different groups annotate, conceptualise, or visualise, and collect additional empirical input on the various research questions of the task. These experiments were to confront humanists with the semantic annotation approach of Pundit. Since the usability of Pundit had been evaluated during the Wittgenstein Incubator, and Pundit and its component had reached a stable stage, experiments appeared to be the best way to proceed.

In parallel, we began working on the "reasoning" aspect of the task. Here, we also started with desk research and then used the preparation of the experiments to include use cases aiming at investigating the actual reasoning of humanists in the context of Linked Data. The report on these aspects of the experiments are provided in the section "Report on Reasoning".

[1] http://dm2e.eu/dhab/

There are several publications and presentations related to Task 3.4. The major contributions are contained as research reports on the respective parts of the task. The first of the research reports is concerned with the Scholarly Domain Model (cf. section 2). An early version of the Scholarly Domain Model has been submitted and published as an extended abstract at the Digital Humanities 2013 conference.[2] The report on the Scholarly Domain Model has been resubmitted to Digital Scholarship in the Humanities (DSH), formerly known as Literary and Linguistic Computing (LLC), and has only slightly been revised and supplemented for this Deliverable. An extended abstract of the report on Reasoning (cf. section 6) has been submitted to the STRiX conference[3] and a full paper will be submitted to DSH in the near future. The report on the experiments (cf. section 5) will be published as separate contributions in association with the respective researchers the experiments have been conducted with. Numerous presentations and talks have been given on the topic of mostly the Scholarly Domain Model. For a list of all publications and presentations given with relation to Task 3.4 confer the Appendix: Related Publications and Presentations.

First we will introduce and discuss the Scholarly Domain Model since it is the basis for the following work. Then, interviews will be discussed which have been conducted in order to collect empirical evidence on the adequacy of the SDM. The section on the development of Pundit will discuss how Pundit has evolved in the light of the feedback of humanists over the course of the project, including a short analysis which Scholarly Activities can be found in Pundit and related components. The section on the experiments with Pundit will report on the outcomes of three workshops which have been conducted in order to collect additional input on the question what humanists can do with Pundit and Linked Data. The last section will report on experiments which have been conducted in order to find out about the kinds of reasoning humanists want to apply to Linked Data. The conclusion summarises the most important findings and gives a few recommendations for future work.

Each major section of the Deliverable can be read separately since the work presented in each section constitutes a coherent sub-part of the Task.

[2] http://dh2013.unl.edu/

[3] http://spraakbanken.gu.se/eng/strix2014

2 Scholarly Domain Model

In this section we will examine how the modelling of research processes in the humanities can inform the development of digital tools created for the enhancement and augmentation of scholarship. In particular, we will focus on how better models of the way in which students and scholars conduct research can be used to support the development of tools that enable users to interact with collections of texts and metadata - including transcription, translation, annotation, and curation.

2.1 Introduction

Over the last decades, the international institutions of research funding have been taking part in a process that could be described as the transition into the digital age. In this respect, they have encouraged a variety of projects for the advancement of the "Digital Humanities",[4] focussing on attempts to further the development of infrastructures for digital scholarship in the humanities. In Europe, for example, the European Strategy Forum on Research Infrastructures (ESFRI)[5] has funded several infrastructure projects such as the Digital Research Infrastructures for the Arts and Humanities (DARIAH)[6] and the Common Language Resources and Technology Infrastructure (CLARIN)[7], which have since been brought together by the Data Service Infrastructure for the Social Sciences and Humanities (DASISH)[8]. Each of these infrastructure projects have, in turn, influenced a number of others on the national level. Apart from the technical requirements of digital information and communication technology, they all have in common the desire to provide the building blocks for a sustainable "Virtual Research Environment" (VRE).[9]

Achieving a constellation of building blocks that is favourable to increasing sustainability is still a major challenge. This is due to many reasons,[10] among them a deficit of systematic investigation into, and hence a deficit in addressing, the actual research practices of humanists and their sustainable representation in the digital realm. For VREs, it is essential to understand the entire scholarly research process and offer applications and services which can support the corresponding workflow.

In this context, the research gap we identified and address is the lack of a model which stresses the importance of creating a bridge connecting the analogue and digital scholarly practices and, most importantly, stresses the recursive relationship between these scholarly practices and the models and applications reflecting on them. This kind of research falls within what is typically called "Digital Humanities" and which we understand as a community of practices, regardless of their particular materiality. We therefore believe that in order to be able to discuss the "Digital Humanities" in a way that goes beyond simply discussing infrastructure so that the aforementioned challenge can be overcome, we need to start from a "modelling process" that allows for the systematic and theoretically grounded building of bridges between practices of humanist research approaches in both the analogue and digital

[4] We understand this term to be grounded in the translation of the German word for "Geisteswissenschaftler" and not in the political sense. Cf. for that matter Gold (2012) and Terras et al. (2013).
[5] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri
[6] http://www.dariah.eu/
[7] http://www.clarin.eu/
[8] http://dasish.eu/about_dasish/
[9] Cf. Candela et al. 2013, respective Research Infrastructure.
[10] Cf. the comprehensive work of the European Science Foundation as presented in Moulin et al. 2011a and b.

world.[11] In this deliverable, we discuss this undertaking and propose a multi-layered model that exemplifies the constituents of our modelling endeavour, which we have labelled the Scholarly Domain Model (SDM).

The SDM has been devised based on the assumption that understanding what John Unsworth (2000) had originally proposed in terms of Scholarly Primitives is central to any such approach at modelling the digital scholarly domain. Unsworth's Primitives are understood as "basic functions common to scholarly activity across disciplines, over time, and independent of theoretical orientation" (Unsworth 2000). Like other models since, the SDM takes up the notion of Primitives and develops them further. Based on analysing and observing the practices of digital scholarship, we are endeavouring to acquire a better understanding of the requirements for instructing the development of sustainable infrastructures that enable scholars to harness the potential of digital technology and hence to develop appropriate digital methodologies. This requires to proceed beyond the establishment of static models to the iterative and continuous activity of "modelling".[12] For this reason, the SDM is conceived as an explicit but not definite set of the constituents of the domain of digital scholarship in the humanities. Similar to Manfred Thaller in his talk "Praising Imperfection" (cf. Thaller 2013), we believe that modelling is the goal, not the model.

In this regard, *Linked Data* standards[13] such as the Resource Description Framework (RDF), Resource Description Framework Schema (RDFS), and Web Ontology Language (OWL) constitute a well suited means for the development of the SDM, because they allow the process of modelling to be iterative and continuous since the graph of semantic statements created is extensible. As we will see, this is also an instance of a still uncommon and emerging way to think of Linked Data as an art with epistemological implications for the practice of modelling the domain of digital scholarship in the humanities (cf. Oldman et al. n.d.).

One of the main activities of the DM2E project has been working on further developing a digital humanities collaboration environment which is built around the semantic annotation tool Pundit originating from the SemLib[14] project. Pundit along with additional modules enables scholars to work with digitised manuscripts in the Linked Open Data (LOD)[15] Web. The development of this collaborative research environment and the modelling process of the SDM have partly informed each other. The results of DM2E are intended to contribute to the emerging digital, networked and distributed environments, well beyond traditional working paradigms in the scholarly culture of the humanities. The SDM plays a pivotal role in this respect as a framework for better understanding scholarly research practices and the ways digital working modes might evolve in the future.

Starting from the Scholarly Primitives by John Unsworth (2000), the SDM was further constructed and refined by analysing the research literature and related models, which will be discussed in the following section. Furthermore, the conceptual input has been subsequently revised and supplemented by empirical evidence collected through a series of interviews with scholars and researchers from the humanities, and experiments using the Linked Data annotation environment Pundit.[16] Finally, the work on the SDM has continuously

[11] Cf. McCarty (2005) as well as Beynon et al. (2006) which delineates our theoretical background of the modelling process.

[12] McCarty (2004), McCarty (2005) as well as Beynon et al. (2006) cf. further section 2 on that matter.

[13] Cf. for the following standards http://www.w3.org/standards/techs/rdf#w3c_all and http://www.w3.org/standards/techs/owl#w3c_all.

[14] http://www.semlibproject.eu/

[15] http://www.w3.org/standards/semanticweb/data

[16] Detailed reports on the interviews and experiments can be found in this Deliverable.

been monitored and counselled by the Digital Humanities Advisory Board (DHAB) where DM2E has brought together scholars of the digital humanities in Europe.[17]

In the following section 2.2, we further motivate our research and discuss the wider context of related research on which we built our model. Section 2.3 offers a detailed description of our proposed Scholarly Domain Model. Section 2.4 provides an outlook on how the model and the modelling could facilitate and support the development of sustainable VREs for scholarship in the humanities.

2.2 From Infrastructure to Modelling the Scholarly Domain

First, we will introduce the wider research context of our work on the Scholarly Domain Model starting with the observation of the predominant focus on infrastructure in a lot of digital humanities projects. Then we will present related research literature and similar modelling efforts.

Infrastructure[18] is required in order to enable advanced collaborative approaches of scholarly work in digital and network based environments. Thus, attempts currently under way to make such infrastructures available are essential, as described by Rockwell (2010) from a North American perspective. Most of these efforts have their roots in the National Science Foundation (NSF) initiative[19] that led to the foundational "Atkins-Report" (cf. Atkins et al. 2003). This report introduced a layered vision of the way technical research infrastructures are related to each other (cf. figure 1).

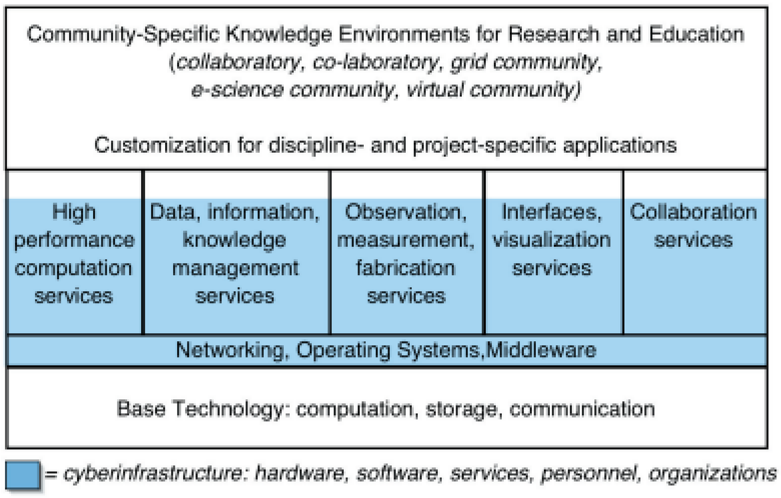

Community-Specific Knowledge Environments for Research and Education
(collaboratory, co-laboratory, grid community,
e-science community, virtual community)
Customization for discipline- and project-specific applications
High performance computation services
Data, information, knowledge management services
Observation, measurement, fabrication services
Interfaces, visualization services
Collaboration services
Networking, Operating Systems,Middleware
Base Technology: computation, storage, communication
🟦 = cyberinfrastructure: hardware, software, services, personnel, organizations

Figure 1. The layered vision to technical research infrastructure from the Atkins-Report.

This "mother of all eScience layer cakes" introduced the hitherto canonical division between the blue area of supporting cyberinfrastructure and the white area of discipline-specific applications. Most initiatives following this report were to focus more or less exclusively on the cyberinfrastructure layer[20] such as the report on "Our Cultural Commonwealth" (Unsworth et al. 2006). The model of thought introduced by this report has also been

[17] http://dm2e.eu/dhab/
[18] Cf. Atkins et al. 2003.
[19] http://www.nsf.gov/
[20] Some passages of the report read as if the "Base Technology" layer was also part of the cyberinfrastructure. And some participants of the group moderated by Dan Atkins may even have wished to place the focus of cyberinfrastructure rather in this base technology area - but this does not invalidate the point made here regarding the division of cyberinfrastructure and the discipline specific application area.

adopted in Europe such as with the e-Science initiative[21] in the UK or the German D-Grid[22] initiative.

An important exception to this exclusively infrastructure driven position was the Bamboo project[23] which included work well beyond the mere building of infrastructure. For instance, the Bamboo project delivered a report on scholarly practices (Bamboo 2010) derived from extensive workshops, an approach that we have partially applied in our own research. Other European and national initiatives initiated a shift from exclusive infrastructure driven positions to content-based focus in the digital humanities. For example, Europeana,[24] as an attempt to make representations of massive amounts of cultural artefacts available on the Web, certainly focuses much more on content than infrastructure, similar to the French humanities research platform Isidore.[25] Still, despite these exceptions, the overall tendency even in the European initiatives is mostly centred on infrastructure.

Infrastructure is not sufficient in itself if we really want to provide the tools and services the researcher needs and will use in the digital, network based environment of the Web, and, in the long-run, want to step beyond emulating traditional scholarly practices. Rockwell (2010) in his section on the “Dangers of Infrastructure” pointed out that, when building an infrastructure, we need to be aware of two major pitfalls: Neither are research infrastructures research “just as roads are not economic activity”, nor should research infrastructures become an end in themselves, where “to sustain infrastructure there develops a class of people whose jobs are tied to infrastructure investment.”

Quite some research has been contributed on the issue of formalising Scholarly Activities over the past decades. Here, we do not present an exhaustive or even extensive review but only a small selection of some of the more recent and essential literature about Scholarly Primitives and related concepts. The SDM has been created starting from and based on this selection.

John Unsworth (2000) conceptualised the Scholarly Primitives as basic functions which are common to any scholarly practice in the humanities independent of discipline, theoretical orientation, or era. He suggested seven recursive and interrelated Scholarly Primitives - discovering, annotating, comparing, referring, sampling, illustrating, and representing - which he saw as the basis for tool-building enterprises for the digital humanities. Since then, Unsworth’s Scholarly Primitives have been often utilised and further revised. And as John Unsworth acknowledged in an interview almost a decade later, his list of scholarly Primitives is certainly not definitive (cf. Unsworth and Tupman 2012). Subsequent research shows that there is no agreement on the exact definition or scope of Scholarly Primitives. However, the approach of using Scholarly Primitives or similar concepts appears to be a valuable and accepted means of structuring and conceptualising the scholarly domain or aspects of it. Therefore we decided to use Unsworth’s conceptualisation of the Scholarly Primitives as a starting point for our own Scholarly Domain Model.

In their “activity centric approach” Palmer et al. (2009) revised Unsworth’s rather static notion of Scholarly Primitives by grouping them into “scholarly information activities”. This approach stresses the vivid character of research and the role of information in the scholarly domain where Primitives form the basic building blocks of larger scholarly information activities. Based on an extensive literature review Palmer et al. identified five core scholarly information activities - searching, collecting, reading, writing, and collaborating - each of

[21] Cf., for example, a Humanities and Arts perspective on the e-Science initiative in the UK in Blanke and Dunn (2006).
[22] http://www.d-grid.de/ as well as, for example, http://www.textgrid.de/.
[23] http://www.projectbamboo.org/
[24] http://www.europeana.eu
[25] http://www.rechercheisidore.fr/

them containing several more granular Primitives, some of them being “cross-cutting”, which means they can be applied to any Scholarly Activity. Furthermore, this study indicates that the “Scholarly Primitives and Activities” exist universally in both the “sciences” and the “humanities” although in different weighting. We took a similar approach, however, in our model the Scholarly Primitives are specialised into more granular Scholarly Activities. Also, while Palmer et al. (2009) only mention different kinds of “stages of a research project”, we embedded Primitives and Activities into a wider context of a process model for research activities.

Whereas Palmer et al. (2009) based their work on an extensive literature review, Brockman et al. (2001: 4) conducted early empirical research on how “humanities scholars think about, organise, and perform their research” and the ramifications for tool building enterprises. Their study suggests four general and intertwined categories of activity: Reading, networking, researching, and writing. They conclude that such analysis of the humanist’s research process constitute essential input to the development of digital tools for the humanities. In 2010, the Bamboo project[26] performed a series of workshops with practitioners from the digital humanities in order to examine scholarly practices. They mapped their findings to the ones of Unsworth (2000) and Palmer et al. (2009). Their aim was to provide a conceptual framework for tool-building enterprises in the digital humanities. The “Scholarly Practice Report” (Bamboo 2010) and the recordings from these workshops are a rich source which helped us to devise the initial Scholarly Primitives for the SDM.

Apart from the research strand opened up by John Unsworth another relevant perspective is provided by the notion of the “methodological commons” introduced by Anderson et al. (2010), on the basis of McCarty and Short (2002). They sketched out an intellectual map which is meant to be a vivid means for mapping out the field of digital humanities. This map is intended to provide a starting point for a framework which may visualise the complex interrelations and interactions between the different disciplines, source materials, methods and technologies involved in scholarly practice of the digital humanities. In the end, and very similar to our conception of the process of modelling, the activity of mapping out the field of digital humanities has to be thought of as a continual process that is the point meant to spark off debate and to ever evolve the diagram further. Anderson et al. (2010) combined the methodological commons with the Scholarly Primitives in order to create a conceptual framework for a tool-building enterprise for the digital humanities in DARIAH. They also stress that those Scholarly Primitives should be extended beyond textual content and consider the Primitives mainly as a means of communication and explanation what traditional research activities digital tools actually enable. Similar to the methodological commons we aim at creating an integrative and vivid model of the research process but focus on its common and discrete functions and its social aspects.

Benardou et al. (2010) probably came closest to our intentions. As part of the DARIAH-EU preparatory work they devised a conceptual model of scholarly research activity which is expressed in terms of the CIDOC CRM.[27] They do not propose a comprehensive list of Scholarly Primitives or Scholarly Activities but, building upon and extending the CRM’s notion of activity, show how scholarly primitives could be operationalised as properties connecting research activities with information objects and propositions, i.e. including argumentation structures. Their proposal goes beyond being a framework for categorising tools but also aims at capturing results from empirical research on scholarly research activity (also cf. Oldman et al. n.d.).

[26] http://www.projectbamboo.org/

[27] http://www.cidoc-crm.org/

The Scholarly Primitives and especially the Scholarly Activities are primarily based on the work by Unsworth (2000), Palmer et al. (2009), and the Bamboo Project (2010). Whereas we started our research on this basis, utilising concepts of Activities, as well as their respective terminology and description, the concepts as they have been included in the “Appendix: Scholarly Activities” have been subsequently appropriated for the Scholarly Domain Model. As a result of this process, some of the Activities have been substantiated, eliminated, revised and renamed or inherited as they were. We emphasise that this list attempts to be explicit but not definite and demands to be further appropriated for future application.

In the past few years, several different approaches to classifying tools and methods have emerged, some sharing the same aim as the SDM, some concentrating on being registries of existing tools. Figure 2 shows the interrelation between the different endeavours.

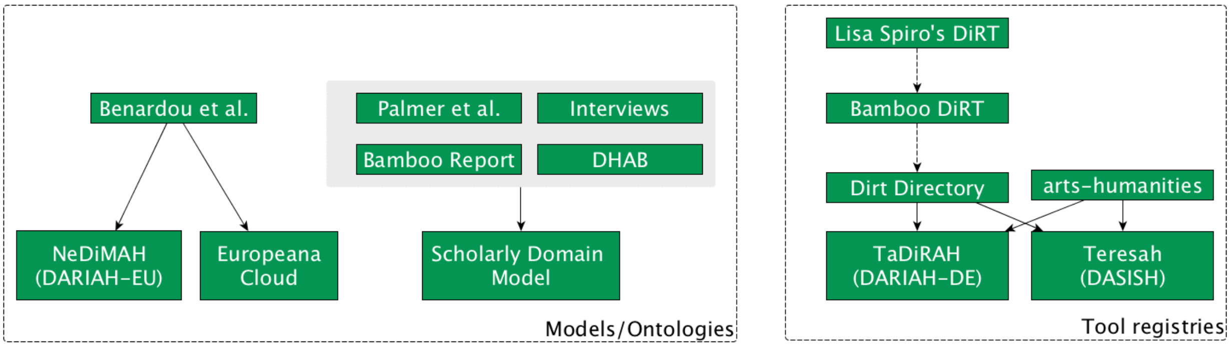


Figure 2. Related Models.

Although being mentioned as an inspirational source in most of the recent literature, there is no taxonomy that is directly derived from Unsworth (2000). However, it serves as an anchor for all the projects mentioned here. Figure 2 shows the genesis of currently active projects, demonstrating that there is a difference in the aims of the taxonomies. Some consider themselves to be mere tool registries while others, like the Network for Digital Methods in the Arts and Humanities (NeDiMah)[28] and the SDM aim to describe the scholarly research practices as a whole.

The Digital Humanities Taxonomy Group[29] develops the Taxonomy of Digital Research Activities in the Humanities (TaDiRAH). The rationale, as quoted from the group’s GitHub account[30], is to help endeavours to collect information on DH tools and methods. The top categories are modelled after the phases of a prototypical research process and contain more specific methods. It is pointed out that the taxonomy is not meant to cover all the methods that might exist in DH, but concentrates on a set of methods that are widely used. In addition to the category of activities there are two lists: techniques and objects. Techniques[31] (e.g. “Brainstorming”, “Searching”, “Encoding”) specify how an activity (e.g. “Visualization”) is actually performed while the list of Objects[32] (e.g. “Metadata”, “Persons”) is a list of objects that the technique can be applied to. Both these lists are open and might change over time. From a theoretical background, the taxonomy bases itself on the concept of Unsworth’s Scholarly Primitives, as well as “the idea of a multi-stage scholarly workflow or research lifecycle”. Also, the taxonomy separates research activities from research objects.

[28] http://www.nedimah.eu/
[29] https://github.com/dhtaxonomy
[30] https://github.com/dhtaxonomy/TaDiRAH/blob/master/introduction.md
[31] https://github.com/dhtaxonomy/TaDiRAH/blob/master/reading/techniques.md
[32] https://github.com/dhtaxonomy/TaDiRAH/blob/master/reading/objects.md

Another tool registry that is currently being developed is the DASISH Tools Registry, or TERESAH (Tools E-Registry for E-Social science, Arts and Humanities).[33] This will remain a pure tool registry and ingested data will come from both arts-humanities.net[34] and Bamboo DiRT.[35]

In contrast, NeDiMAH strives to build a formal ontology for the digital humanities including a classification and a shared vocabulary.[36] It is still in active development by DARIAH-EU's VCC2. In a presentation given at Luxembourg DH conference, December 5th 2012, Lorna Hughes stressed the usefulness of the project as it would "formalize and codify the expression of work in DH",[37] meaning also that an endeavour like this could help and produce a common nomenclature in the digital humanities and that the use of DH methods would gain a greater academic credibility when being grounded on a theoretical basis.[38] NeDiMAH and TaDiRAH are two closely related projects, coming from the DARIAH-EU and DARIAH-DE[39] contexts respectively. Whereas TaDiRAH has a very practical approach, NeDiMAH tries to address the formalisation and classification of methods in the field. It is planned, however, to integrate TaDiRAH into NeDiMAH at a later point. The efforts in Europeana Cloud[40] are also linked to DARIAH-EU's VCC2 "Research and Education Liaison" and through this to NeDiMAH.[41] One of its ambitions is to contribute to the future Europeana Research platform. The most important report here is Deliverable 1.2 of the Europeana Cloud project (cf. Benardou et al. 2013), a desk research on the current situation of digital research practices, tools and scholarly content which gives an extensive overview over current and past studies. They conclude that even though the use of digital sources and tools has become more common and that methods in the digital humanities reflect on this phenomenon, there is still the need to support the building of infrastructures by more research on the way scholars of the humanities interact with the digital domain.

The literature and models we presented above provided us with valuable input for important categories and the overall design of our modelling approach. However, we found that these models lack a perspective that we consider important for our purposes: Although humanities exhibit an increasing drift into the digital, the major part of the scientific community is not using dedicated digital humanities tools. Rather, scholars rely on well-known but not necessary the best suited software. Thus, maintaining tool taxonomies and classifications of digital methods are necessary, but not sufficient steps on the way to modelling and supporting scholarly work as a whole.

The SDM which we are going to present next proposes a non-static model whose constituents cover the analogue, traditional activities of the humanities and put them into a general, integrative model of research that also considers the digital context.

[33] http://teresah.angular.dev.dasish.eu/#/
[34] http://www.arts-humanities.net/
[35] http://dirtdirectory.org/
[36] http://www.nedimah.eu/workgroups/development-ict-methods-taxonomy
[37] Cf. Hughes (2013).
[38] Further information can be found in the minutes from a workshop in 2013: http://www.nedimah.eu/reports/scholarly-practices-research-and-methods-ontology-workshop-methods-taxonomy-workshop-iii.
[39] https://de.dariah.eu/
[40] http://pro.europeana.eu/web/europeana-cloud
[41] https://dariah.eu/activities/research-and-education.html

2.3 The Scholarly Domain Model

The model we are proposing consists of four different layers of abstraction which will be described in the following sections. These layers are Areas, Scholarly Primitives, Scholarly Activities and Scholarly Operations.

The *Areas* represent the general stages of scholarly work, whose central point is of course research, but which also covers aspects of a circular workflow and surrounding, contextual Activities like social and administrative aspects which influence the research process. The next layer consists of the *Scholarly Primitives* that form the most abstract description of scholarly practices in the model. The Primitives are located mainly in the Research Area, but extend also in other Areas. The motivation for these Primitives is what we think the simplest description of the research process: Interpretative Modelling, Exploration, Aggregation, Augmentation and Externalisation. The third layer consists of *Scholarly Activities*, a set of categories for describing possible research processes. The categories we propose are still generic and not domain-specific constituents. In contrast to the Primitives, the Activities refer to particular yet generic parts of the research process which, in principle, may occur in any sequence or constellation. Thus Activities do not have an exclusive or definite subclass relation to Primitives but may be seen as relating to one or more Primitives based on the particular context of their application. The *Scholarly Operations* form the most concrete layer of the model. On this level, the Activities are viewed through the lens of a specific application scenario, i.e. including other constituents of the SDM such as the Actors that perform Scholarly Activities within their respective research process, determined by a Social Context as well as the applications and material at hand.

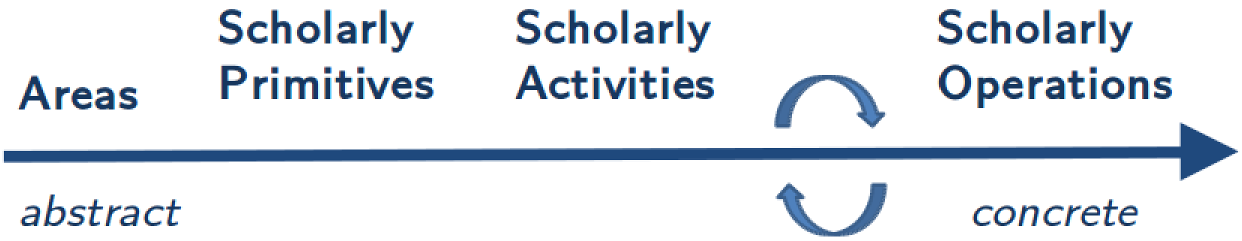


Figure 3. Layers of Abstraction.

We chose this layered division of the model (cf. figure 3), in order to be able to adapt the model to a number of possible applications during the modelling process. With this framework, the scholarly domain can be modelled on four different levels where the first three provide a systematic and structured vocabulary for the analysis of the scholarly domain while the fourth one is concerned with their respective observation in practice. The potential benefit of this is that the model, in particular seen as a modelling process, is better suited to react to the requirements of the continuous development of research infrastructures as well as of the scholars, those infrastructures are developed for. Recursion, or the facility to integrate what is “missing” is crucial for the model to be easy to apply as well as to adapt. This recursive and adaptive modelling process can be driven by the use (as of now) of RDF to make components interact with each other on data level.

The focus of the representation within the SDM, described in the next section, is on the scholarly practices represented in the four layers of abstraction. Other pivotal constituents of the domain, such as the scholars, the Actor, or the representations of the objects of their research are not included.

2.3.1 Areas

The five proposed Areas - Input, Research, Output, Documentation and Social Context - form the uppermost and integral part of the Scholarly Domain Model with the central Area being Research, as shown in figure 4. We have chosen to start here, because it reflects the cycle of scholarly work with its different phases of gathering input for research, the act of dealing with the input and the externalisation of results. Hence the two additional Areas Input and Output. Furthermore, we cannot examine the research process in isolation, but need to address its context too, in particular the Social Context and Documentation.

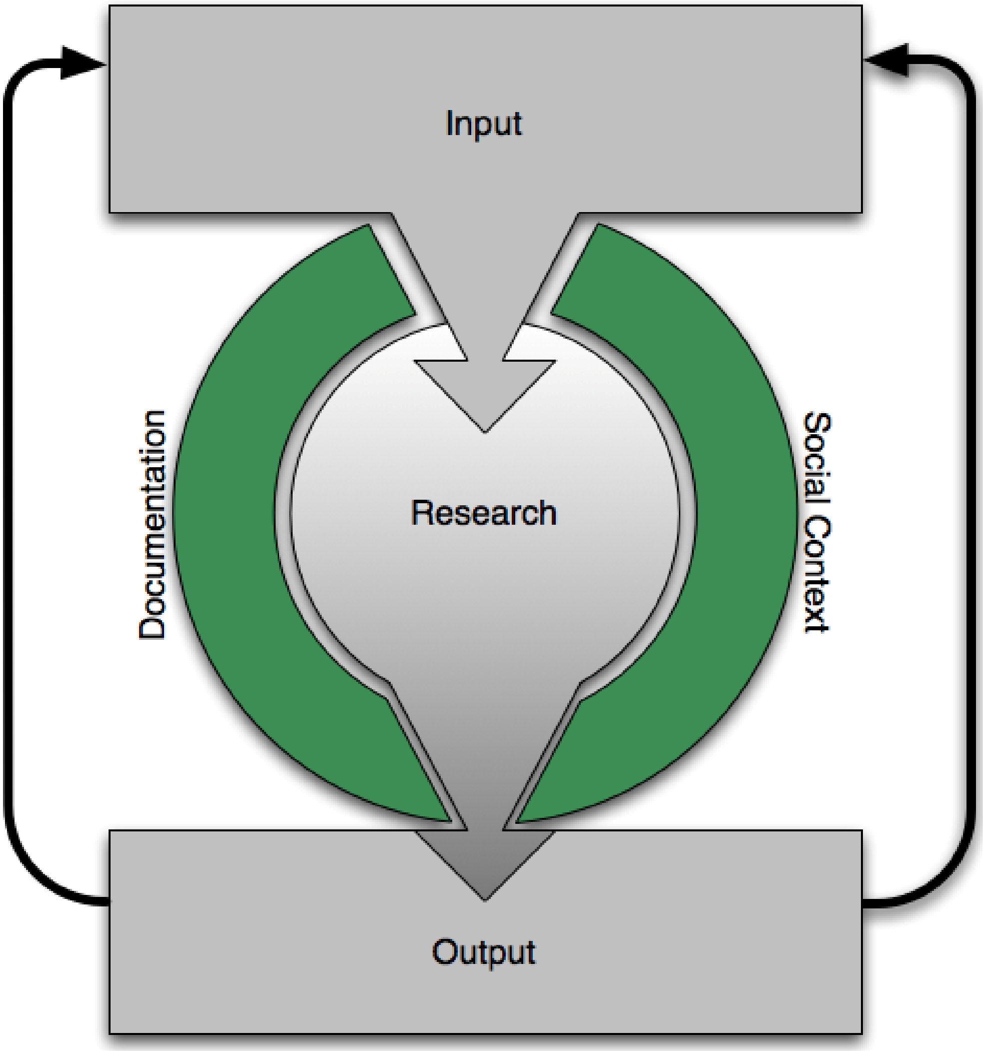


Figure 4. The five Areas of the SDM.

The arrows in the figure imply a sequential grouping where one Area probes into the next one. The input flows into research, research manifests and is being condensed in that process as output. The output, either intermediate results or final results, then serves again as input in another iteration of the research process.

Input covers all Activities and objects that deal with the exploration and aggregation of material that will be used for research. For example, Activities, that range from Searching the Web or Browsing library shelves, and the excavation at an archaeological site to the Selection and Assessment of objects relevant for the Research.

The arrow of input protruding into the Area of research shows that these processes of exploration and building of the corpus could already be considered to be parts of the research Area. The term *corpus* in our model denotes any information object the scholar collects or otherwise aggregates for the purposes of research including personal collections of research data. It is the representation of the objects of research in the SDM. The research process is not linear which leads to iterative modifications of the corpus. For example, in a later stage, when doing research properly speaking, a scholar might discover that elements are missing from her corpus which she then needs to adjust by going back into a phase of exploration.

Two additional sources serve as input to the research process properly speaking as they exist prior to this research: *Referential Data* and *Referential Structures*, both of which are explicitly shown only in figure 4. Referential data are, for example, dictionaries that would be useful for someone reading an 18th century political tract, and wanting to see other contexts from this period in which a certain word (e. g. “liberty”) is used. Another example are Semantic Web ontologies and Linked Data resources as for instance used in the Research Space project[42] or in the Isidore environment.[43] On the other hand, referential structures such as grammar type resources, rule systems and others pre-exist the actual research but here again are placed in the interfacing area since – as we will see below – interaction with these in the sense of corpus contextualisation is one of the first steps in research.

Before considering the research Area in detail, the core of this layer of the SDM, we will first have a look at the *output* Area as well as the contextual Areas, social context and documentation.

Leaving the central Area of research as a black box for the time being, final and intermediate results of research are shared and disseminated as *Output*. Information that has been refined during research is now being externalised as a stable and citable information object, irrespective of its material carrier that becomes subject to reference for either private use limited sharing within groups, or general publication. The potential of this externalisation to enter subsequent iterations of the research cycle is assessed. Output typically entails also a change in availability of these research results: what has been kept in seclusion until now or has been shared with only a few colleagues and members of working groups is released to the public.

Note that – as was also the case with Input – there is overlap with research properly speaking. The discursive and technical organisation of research output is to some extent determined by the way it will be published at a later stage and vice versa. And this overlap may be significantly larger in humanities scholarship as compared to the so called empirical sciences: As was shown in Gradmann and Meister (2008), research and results in the empirical sciences can be considered to be totally disjoint, as in the case of an experiment and the paper reporting on it, whereas in the humanities there is a tendency for publication format, research corpora and scholarly discourse to be highly intertwined.

Furthermore, the bulk of the output may well come from research, but the social context and the documentation can certainly be considered relevant sources of output as in the cases of published citation analysis or project reports. Thus, the central Area, research, is additionally highlighted by the two remaining Areas, documentation and social context, depicted in green in figure 4. These two Areas form the context in which research is embedded.

[42] http://www.researchspace.org/
[43] http://www.rechercheisidore.fr/

The Area of *Documentation* reflects on the fact that research involves the externalisation of a form of meta-discourse to create accountability, transparency and the ability to retrace the single steps of research.

This may include informal exchanges related to research progress and also formal reports that need to be given to funding agencies. Taking the digital humanities as an example, correspondence via email about research, the keeping of notebooks and comments made when checking in source code into version control is an important form of documentation.

Also, it facilitates the interpretation of research processes and results over time in creating a narrative context – which itself can become subject of research. Furthermore, there is a need for a discourse about research itself in Science Studies as well as in the history of science, and documentation provides the material basis for this discourse.

The Area of *Social Context* reflects on the fact that research is determined by the socio-historical situation in which it occurs. This includes such factors as domain specific research practices, the customs of research communities as well as national and international academic cultures. The SDM accounts for this by acknowledging the existence of the *Social Context* and that it affects the Scholarly Activities and Scholarly Operations carried out by researchers. A fact that is overlooked easily by ongoing infrastructure projects: social influences such as research practices are important, especially in interdisciplinary endeavours. And, as already stated for documentation, it can inform the meta-discourse about the research process. Rules, control and incentives are key notions in this social context of VREs.

The importance of including the *Social Context* in the SDM can be made clear by the example of citation which determines the way a citation looks like in a publication such as the conventions in a discipline or style guides by publishers, who is actually cited, often caused by political aspects or regarding the career of the author, and what is cited.

Research is not as exclusively content oriented and content driven as many of us tend to think. Many aspects are often motivated or constrained by the social context, leading to research results being a complex amalgam of content and its apprehension by the scholarly community. Collaboration with others in the research process is a sensible issue in this respect, requiring highly granular and controllable data privacy settings. What is secluded in one moment may be shared with a close community in a second moment and after publication inversely would require extreme visibility in order to obtain references, citations and crediting by the scholarly community.

The *Research* Area centres on those elements which constitute scholarly research at its core. The other Areas – Input, Output, Documentation and Social Context – share several of their elements and interact with the constituents and Activities in the research Area. We will first describe how the various Areas interact with the research Area.

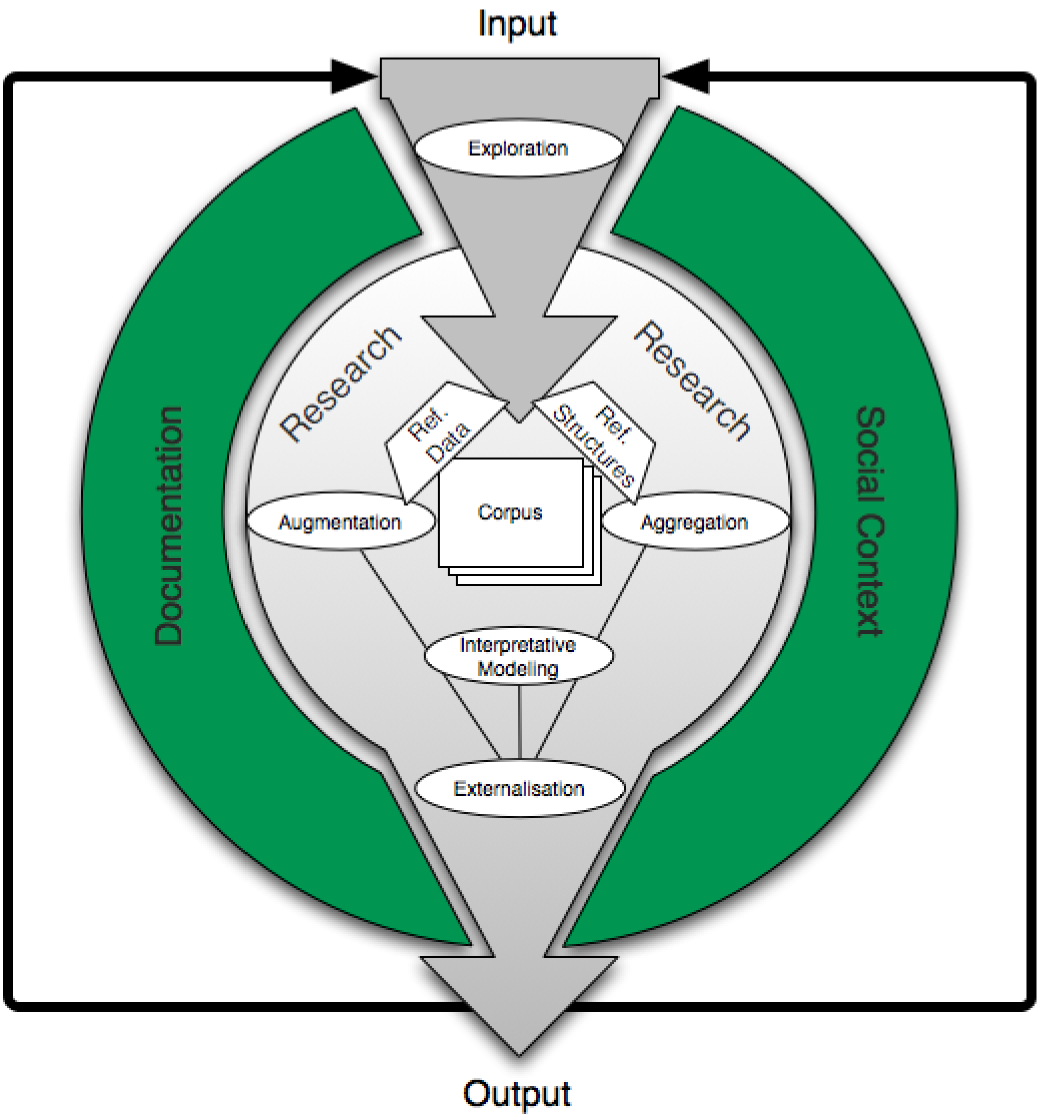


Figure 5. The Research Area of the SDM.

Apart from the Primitives present on this layer there are three additional constituents in the research Area. The most important of those is the corpus, the body of sources the scholar decides to work with. In this abstract model we refrain from specifying anything particular about the consistence of the corpus, but it might contain any sort of information objects, and any sort of data that are manageable by machines including their metadata and data model. Adding objects to and removing from the corpus might be as simple as bookmarking a page in the Web browser or returning a book to the library or as hard as excavating the ruins of a Roman temple. The other two are, as already mentioned above, referential data and referential structures. They are auxiliary entities which are used to contextualise elements of the corpus, for example linking to authority files or Linked Data resources, or to embed one's own research into a broader context, like a theoretical framework. This shows that the various Areas are not strictly separated but are fading into each other.

Within the SDM, the research process begins with creating the corpus and contextualising its elements using referential data and structures. The basic process underlying contextualisation, and for that manner conceptualising, within the framework of the SDM is the Primitive interpretative modelling**,** as it is underlying all scholarly research, representing the process of "understanding" the corpus and its constituents for the purposes of the research process. We assume that any kind of Scholarly Primitive and Scholarly Activity is always grounded in this Primitive. Interpretative modelling forms therefore the core of the research Area. In addition, we propose at least four additional Scholarly Primitives: exploration, aggregation, augmentation and externalisation.

In order to show how the other Areas are related to research, consider the following example of a researcher analysing the works of Ludwig Wittgenstein. This imagined researcher would first of all explore and determine the input of her research and build a corpus of relevant sources and articles, possibly from the Wittgenstein Repository[44] and Wittgenstein Source[45] provided by the Wittgenstein Archives Bergen (WAB),[46] and possibly by utilising a faceted browser, the Wittgenstein Ontology Explorer.[47] The faceted browser helps to iteratively focus and zoom into the sources and metadata, and restrict the corpus to a selection of items relevant for the specific research question, e.g. whether and which visual analogies occur in the context of Wittgenstein's remarks on the nature of philosophy. With the ongoing research process, she would keep her working group, for example other Wittgenstein scholars that have made their Pundit notebooks public, updated on the progress through sharing it with them in her own Pundit notebook (social context). She would document any additional findings in separate Pundit notebooks and would inform her university on the progress of work (documentation). Once there is an accumulation of valid and valuable research results, she would generate output by for example presenting the results at the annual international Wittgenstein Symposium in Kirchberg.[48] Ideally, she would also publish her aggregated set of research data extracted from the corpus – together with the processing methods she had used in her research for others to use as input for subsequent research projects.

This example also makes clear that the picture should not be read as a static arrangement of components nor as their linear succession in the Input (start) → Research → Output (end) sequence. The arrows pointing from the output to input are meant to visualise a circular process in which the output of one iteration can be and typically is input for the next. To really comply with the complexity of the research process its recursive nature might even be organised as a spiral in order to indicate progress instead of eternal repetition.

Next, we will have a closer look at the Scholarly Primitives, which we think represent the most basic constituents of any humanistic research process.

2.3.2 Scholarly Primitives

The *Scholarly Primitives* constitute the most generic and principal parts of any research process in the humanities and form the second most abstract building blocks of the model. They facilitate the initial description of research processes in a very abstract but still generic way and constitute a basis for proceeding to more specific representations.

The basic set of Primitives that we propose are interpretative modelling, exploration, aggregation, augmentation and externalisation. These are inspired by the work of Unsworth (2000), but have been refined further by the study of literature (cf. section 2.2), interviews that we conducted and the counsel of the Digital Humanities Advisory Board (DHAB) as well as observations during experiments with Pundit.

The Primitive *Exploration* is located – as already stated above – primarily in the Input Area and thus happens in a pre- or inter-research state. Exploration is about serendipitously

[44] http://www.wittgensteinrepository.org/
[45] http://www.wittgensteinsource.org/
[46] http://wab.uib.no/
[47] http://purl.org/net/dm2e/wab/search
[48] http://www.alws.at/index.php/symposium

navigating networks of related information objects that will lead to the creation of the corpus. The corpus that is then gradually built up will be the object of Activities like direct searching, browsing and rearranging, but these kinds of tasks are situated on a less abstract level of the model on the level of Scholarly Activities. In this regard, the process of direct searching must be differentiated and can be seen as a particular case of exploration.

As stated above in the section on the Area of research, *Interpretative Modelling* is the basic constituent Primitive that makes up the central element in research and serves as a hub for the other Primitives. The process of “understanding” is what it represents at its core as it revolves around the corpus by contextualising and conceptualising its elements, successively reaggregating and rearranging it to finally be able to externalise ideas are the core Activities here.

The Primitive *Aggregation* consists basically of arranging or rearranging the corpus elements. Filtering and sampling are examples of such aggregation activities that typically result in rearranged elements of the corpus such as, for instance, the pages of a digital edition arranged according to their relevance for the research process or the pieces of a vase found in an excavation arranged for their reconstruction.

Augmentation adds to the elements of the corpus. Annotations and comments are typical examples, but also context links added to the corpus elements. Such augmentations are results of research in their own right, even though their potential for publication is controversy among scholars in the humanities.

Finally, instances of *Externalisation* such as critical texts, textual interpretation or visualisations have to be produced to make the results of interpretative modelling and therefore the research process “readable”.

This list of Scholarly Primitives, and also the list of Scholarly Activities which will be discussed in the following section, are taken from our research undertaken in the context of the DM2E project (cf. section 2.2). Some scholars may find that our Scholarly Primitives do not capture or capture incompletely what they consider to be the Primitives of their own field, or they may feel uneasy with the terminology. In the context of our perspective on modelling we see the SDM as an abstract proposal that provides a domain-independent framework open to further iterations of adaption and specification for the application in more domain specific scenarios. The particular Primitives and Activities are explicit, but not definite.

To continue with our example from the previous section we might take a more specific look at the research progress. In the phase of *exploration*, the Wittgenstein scholar would browse through the Nachlass on Wittgenstein Source and secondary literature in the Wittgenstein Repository, and might also use catalogues and finding aids for building up the corpus such as the Wittgenstein ontology[49] provided by the WAB (aggregation). The corpus is enriched (augmentation) by linking the sources to the Wittgenstein ontology (referential structures) and by looking up references in a lexicon, for example the Glock Wittgenstein Dictionary[50] (referential data), and hereby contributing to augment the original ontology further. The central part of research, the interpretative modelling will take its course. Finally, as an act of externalisation, the results are written up as an article to be submitted to a journal, for example the Open Access Nordic Wittgenstein Review.[51]

[49] http://wab.uib.no/cost-a32_philospace/wittgenstein.owl

[50] Glock (1996).

[51] http://www.nordicwittgensteinreview.com/

2.3.3 Scholarly Activities

The *Scholarly Activities* constitute the most concrete of the abstract layers of the SDM. As with the Primitives, the Activities chosen here reflect the results of the underlying research (cf. section 2.2 above). We propose 25 different Scholarly Activities.[52] We do not consider this list to be definitive, in particular in terms of their number or the labels used for the Activities as well as the scope notes used to describe them. Nevertheless, since many of the Activities on the list are common in the literature as well as in the scholarly work in the humanities, they can be regarded as a recommendation. Despite the fact that a list like this might be subject to further specification for concrete application scenarios, we want to emphasise that the observations made during our research are not exhaustive. We encourage more systematic work on these Primitives, Activities and their ontological formalisation.[53]

Despite the fact the Activities are conceived to be more specific than Primitives, the SDM does not consider their relation to be strict or hierarchical and that it is possible that each Activity can be related to one or more Primitives.[54] The difference between Activities and Primitives can be found in the different layer of abstraction used for the analysis and description of a part of the research process. As previously mentioned, Primitives and Activities typically materialise as sequences, that iterate, or in specific constellations. Therefore, it may appear difficult to determine the relation between Activities and Primitives while observing them as one part of a research process or another. For example, as mentioned earlier, there is some form of interpretative modelling involved in all scholarly research practices.[55] Furthermore Activities and Primitives may also be part of one of the Areas social context, documentation, input, or output.

Since the proposed list of Scholarly Activities, contains 25 items, we refrain from discussing each one individually, but we discuss two Activities as examples: annotating and contextualising. As described in the scope notes (cf. 2.6 Appendix: Scholarly Activities), Annotating is considered to be the Activity of "adding any kind of notes or markings to elements of the corpus". This results in enriching an element of the corpus with additional data, for example, this could be a note written in the margins on the page of a book or – as will be seen below – the mark-up of an electronic resource using RDF triples. The creation of an annotation is accompanied by a series of other Activities. As annotating itself can also be an act of writing, what is being written down can be an act of translating, contextualising or comparing. At the same time, this piece of information is another element that is being added to the corpus. Another important and far-reaching Activity is contextualising which we already encountered earlier. This one would be related to the interpretative modelling Primitive, but connects items of the corpus to referential data and referential structures. Thus, relationships are created either between objects that are part of the corpus, but also between objects in the corpus and external sources. As before with annotating, contextualising also resonates in other Activities. A part of the contextualisation is often a reference or a link to another source, so that in this case referring/linking is an adjunctive Activity.

[52] The complete list of Scholarly Activities together with scope notes can be found in chapter 2.6.

[53] Cf. a draft version of the SDM as an RDFS/OWL ontology can be found at http://webprotege.stanford.edu/#Edit:projectId=32a9b5a3-0781-4846-b195-980482fe54c4.

[54] This is not imperative, cf. for example Palmer et al. (2009), who strictly relate Activities to Primitives. Both modelling practices have their advantages and disadvantages. The stricter the relations are the harder it gets to differentiate the vocabulary further.

[55] The report on reasoning experiments conducted with Pundit discusses an example of how interpretative modelling may materialise and be translated into a digital, Linked Data context.

2.3.4 Scholarly Operations

The *Scholarly Operations* are the concretisation of the Scholarly Activities for a specific application scenario. This concretisation therefore depends on the purpose or the focus of the observation that is intended for the respective scenario. An Activity could be translated into a variety of Operations, with a variety of different constituents, for example regarding citation, as an instance of the Activity referring/linking, the focus could either lie on quantitative aspects of citation behaviour or on qualitative aspects such as different types of citation relations. Thus, for observations of Scholarly Operations focused on the quantitative aspects the actors and a model of their social context, are imperative constituents. Whereas the observation focused on qualitative aspects, might require different constituents regarding the linguistic classification of citations.

In addition to that, as each scholarly discipline or community has its own specific requirements, concerning the applications and the conventions of their scholarly practices, further constituents would have to be determined for the specification of application scenarios. The scholarly practices of the Activity comparing, for example, vary greatly in different disciplines. When comparing two or more different versions of a Middle High German manuscript, the differences between the versions can be computed and visualised by software and might serve as a basis for a critical edition of the manuscript, provided the input texts contain appropriate mark-up. Scholars of Art History or similar disciplines dealing mainly with images will have other needs and means to assess the differences or similarities of the objects in comparison.

Since, the focus of this deliverable on modelling the Scholarly Domain focuses on the description of the process rather than the description of the application of a model, the following does neither attempt to provide a comprehensive description nor to conduct a systematic investigation on “how” the various abstract Scholarly Activities could materialise in concrete Scholarly Operations. Nevertheless the next section does attempt to delineate their particular relation in examples from the context of DM2E. We will discuss how the SDM and its practice of modelling could be used to instruct the development of VREs for digital scholarship in the humanities and how the use of the Resource Description Framework (RDF) as a principal data model could help to sustain its operations.

2.4 From the SDM to Modelling the Scholarly Domain

The current version of the scholarly research platform (Pundit) enables various Scholarly Activities on an application level such as providing facilities for the collection and creation of vocabularies as well as annotations.

For research infrastructures to be able to sustain digital scholarship in the humanities, we believe that the scholarly practices as well as the continuous development of applications by integrating the lessons learned through the observations of user behaviour has to be taken into account. Therefore the SDM has not been devised to be another attempt to establish a static model but rather to instigate an iterative and continuous process of modelling.[56] For this reason, the SDM is conceived to provide an explicit but not definite set of the constituents of the domain of digital scholarship in the humanities.

In DM2E, we conducted a series of experiments[57] in order to approach the observation of the Scholarly Operations within the framework of different application scenarios associated with the Pundit environment. The experiments demonstrate how the manifestations of the

[56] Cf. footnote 12, and further Section 2.2 on that matter.
[57] Cf. full report in this Deliverable.

Scholarly Activity annotating as RDF vocabularies and statements vary in respect to the different research processes of interpretative research in the humanities. The implementation of the Scholarly Activities in applications is a prerequisite for systematic observation of how they specialise into Scholarly Operations in different application scenarios. In this context, the RDF data model which underlies the scholarly research platform developed in DM2E,[58] is a suitable means not only to connect the Activities on a data level, for example, to make annotations explorable alongside the vocabulary used for the annotations, but also to create explicit and formal representations of Scholarly Operations in the first place. The translation of different research interests into simple annotation vocabularies represents one of the possibilities to create Scholarly Operations for observation. They operationalise, i.e. explicate and formalise, the Activity of annotating as RDF statements along with various conventions and guidelines which means that Scholarly Operations may very well consist of different constituents. Furthermore, the experiments also demonstrated that interpretative modelling is indeed influencing and present during the application of the vocabulary, the Activity of annotating, but also during the creation of the vocabulary and the evaluation of the results, in this case through visualisation in faceted browsers. In this context, the terminology of the SDM provides a framework for systematic investigation and operationalisation of scholarly practices, i.e. their translation, again in our case into a Linked Data environment.

The translation of the Scholarly Activity annotating, as well as the translation of the research interest of the respective processes into Scholarly Operations unveiled the inherent relationship between the practice of modelling and the scholarly practices. The Scholarly Operations are mere constructs in the context of specific use cases determined by what we want to observe and what we can observe in particular research processes. As such, Scholarly Operations express and formalise what we would like to and what is possible to analyse and hence may instruct the further development of applications and methodology. In other words, Scholarly Operations are constructs of observation, and as such they serve the purpose of analysis which again serves development of tools and their application.

Secondly, the translation and application of the Scholarly Operations unveils the relationship between modelling and methodological reflexion as the research process is conducted. Scholarly Operations, in the current example the annotation vocabularies, evolve since during their application new constructs may emerge to be represented.

The experiments also suggested that the annotation acts conducted in the context of the interpretative work could be further structured into templates of several combined statements which will be reused. Such RDF-templates are one example of a first step in the direction to substantiate the process of recursion and to be able to approach the representation of the Scholarly Domain by “modelling” rather than by a “model”. Such templates can be modelled in RDF as sets of triples which describe the kinds of statements involved in certain Activities. For example, a template for the Activity selecting may contain a criterion for that Operation, an actor who performs it, as well as the item which is either removed from or added to a corpus and related metadata as constituents. Since RDF allows to specialise properties and classes, communities or single users may create more specific statements within a particular RDF-template. By using such templates, we connect the abstract and conceptual level of the SDM, the model, with the concrete and explicit level of modelling and performed Scholarly Activities.

The second step in the direction to substantiate the process of recursion includes monitoring, either in the analogue or the digital realm. The latter in particular has the potential to proceed to further and to instruct the development of applications through the automatisation of the observation of their usage for their successive analysis to adapt the

[58] Cf. Grassi et al. 2012.

applications according to the actual conduct of scholarly practice. The Patterns identified in such a monitoring of Operations can, for example, be fed back for the adaption of the aforementioned templates and thereby retain the adaptive modelling process. A potential use case has been discussed with the project “Virtual and Real Architecture of Knowledge”,[59] a part of the project “Image, Knowledge, Gestaltung”[60] at Humboldt-Universität zu Berlin,[61] who are planning to monitor and record all digital and analogue interaction of researchers within a laboratory and to extract and model typical patterns of behaviour. The SDM has been taken under consideration to provide an ontological framework for the representation of such patterns since it provides enough flexibility to provide a starting point for such an endeavour. The observed patterns of usage and user behaviour could be integrated into the SDM representation with RDFS/OWL, and consequently be implemented into an application such as Pundit to substantiate the monitoring, for example, of Activities for the documentation of the respective parts of the research process. For the SDM as a framework for integration one of its benefits may become apparent, when it is taken into consideration that the extent of the automated creation of machine-processable data from monitoring Activities also impacts the potential subjects for analysis.

Nevertheless, the experiences we had during the experiments as well as the discussions regarding the potentials of the application of Linked Data and Reasoning technologies in humanities scholarship, as found in Oldman et al. (n.d.), point to the fact, that as recognisable and significant they may be, as careful and delicate they have to be treated not to overestimate future developments. In all described cases, the limits of such RDFS/OWL formalisations need to be identified and kept well in mind in order not to move into the “Artificial Intelligence Rathole”, as adequately termed by Wendy Hall.[62] The aim cannot be to substitute creative thinking, as has been identified in Gradmann (2010), but to assist the scholar during the research process with functionality that, on the one hand, remains rooted in traditional and established processes but, on the other hand, also allows to go beyond using digital infrastructure for the emulation of traditional Scholarly Activity. That is why *modelling* is so important to be thought of as a continual and iterative process that integrates the development of the applications of digital scholarship as well as the basis, in which their use is grounded, the scholarly practices of the humanities.

2.5 Conclusion

In this section we presented the Scholarly Domain Model which has been developed within the context of the DM2E project.

In the light of a recognisable deficit in conceptual work on the constituents of scholarship in the digital humanities and a predominance of infrastructure-oriented projects in the field, the SDM provides a framework for the systematic investigation of the relation between scholarly practices and the emergence of digital practices and methodology in continuously evolving Virtual Research Environments (VRE).

Despite the fact that the SDM has been devised in the context of applications based on Linked Data, it is independent from particular representations and meant to be applicable as a reference model for the discussion, evaluation and development of digital research infrastructures for the humanities. The SDM allows to create representations of the workflow of digital humanists and to function as a terminological bridge between the humanities and digital applications. Only if we better understand how scholars undertake their research now

[59] http://www.interdisciplinary-laboratory.hu-berlin.de/en/Virtual-and-Real-Architecture-of-Knowledge
[60] http://www.interdisciplinary-laboratory.hu-berlin.de/en
[61] https://www.hu-berlin.de/?set_language=en&cl=en
[62] At the Cultural Heritage and the Semantic Web British Museum and UCL Study Day, British Museum, London, January 2011.

and in the past and how their functional framework might be adequately translated to the digital environment, we might actually propagate new digital modes of working. Furthermore, the SDM differs from similar approaches in so far as it approaches the scholarly domain from a more comprehensive perspective and tries to integrate Primitives of the process of scholarly work and various layers of abstraction rather than isolated acts. The model stresses the importance of recursive and continual modelling processes in order to adapt VREs to evolving scholarly practices. Then again, we believe the modelling is the goal, not the model.

Acknowledgments

The following members and associates of the Digital Humanists Advisory Board (DHAB) contributed to the development of the Scholarly Domain Model and the writing of this section:

Tobias Blanke (UK), Sally Chambers (Germany), Alastair Dunning (Netherlands), Stefan Gradmann (Belgium), Jonathan Gray (UK), Gerhard Lauer (Germany), Christian Morbidoni (Italy), Alois Pichler (Norway), Jürgen Renn (Germany), Laurent Romary (Germany/France), Felix Sasaki (Germany), Susan Schreibman (Ireland), Claire Warwick (UK), Dirk Wintergrün (Germany)

2.6 Appendix: Scholarly Activities[63]

(Direct) Searching	Searching with a well-defined goal ("known-item" searches) for specific information or objects of interest which also "involves deciding where and how to look for information" (cf. Palmer et al. (2009): 9-11).
Discovering / Foraging	Discovering objects of interest or information through various aids including conversational means. Foraging stresses the aspect of "discovery" as an alternative to (direct) searching (cf. Bamboo (2010): 3-4).
Browsing	Exploratory and investigative strategy employed to find information in unfamiliar domains or topics. May utilise various exploratory means such as database search, archival aids, and conversation with domain experts or translating unfamiliar terminology (cf. Palmer et al. (2009): 14-15).
Probing	Exploratory and investigative strategy employed to find information in unfamiliar domains or topics. May utilise various exploratory means such as database search, archival aids, and conversation with domain experts or translating unfamiliar terminology.
Chaining	Following chains of citations or references either performed as backward chaining (footnote chasing, following references) or forward chaining (citation searching) (cf. Palmer et al. (2009): 11-13).
Monitoring	Keeping constantly and periodically track of developments and news in a field or related to a topic. Essentially an exploratory Activity which might entail other Activities such as chaining, searching, browsing, scanning and reading (cf. Palmer et al. (2009): 29-30).
Reading	Close reading, but might include other kinds and stages of reading such as scanning or systematic skimming, prior to close reading or rereading (cf. Palmer et al. (2009): 19-21).
Contextualising / Conceptualising	Adding to the corpus referential structures or referential data by creating relationships between one and more of it elements. Can be seen as more special type of referring/linking (cf. Bamboo (2010): 5-6).
Translating	Converting and interpreting of new terminology, concepts, theories, methods, etc. for oneself but also for different audiences (cf. Palmer et al. (2009): 31).
Assessing	Determining the quality of an object of interest or information in terms of its relevance, utility, provenance etc. (cf. Palmer et al. (2009): 20-21).
Comparing	Measuring the differences between elements in terms of their structural and conceptual features (cf. Unsworth (2000)).
Synthesising / Filtering	Synthesising / Filtering Generating a (temporary) view on the corpus on the basis of one or more criteria. Can also be part of the exploration process (cf. Bamboo (2010): 4-5).

[63] Where concepts of Scholarly Activities have been essentially reused or remain close the original conceptualisation, references to the original and closest descriptions are provided. The formal scope notes provided here are mostly more exclusive than the original descriptions and reflect our particular interpretation and conceptualisation of the original concepts. Where no reference is given the Scholarly Activity has no appropriate equivalent.

Sampling	Sampling is a specific subtype of selection in so far as it constitutes a new corpus (the sample) as a subset of the original corpus. Both Selecting and Sampling re-arrange a corpus into a new state or constitute a new one. However, sampling is always performed on an existing corpus (cf. Unsworth (2000)).
Organising	Applying or devising (personal) organisational systems and tools for storing and managing the corpus, its contents or other collections (cf. Palmer et al. (2009): 18-19).
Collecting / Gathering	Building (personal) collections for current or long-term research including any kind of objects of interest and information (cf. Palmer et al. (2009): 16-18).
Referring / Linking	Referencing or linking between two elements, e.g. via a hypertext link or by making a citation (cf. Unsworth (2000)).
Annotating	Adding any kind of notes or markings to any part or element of the corpus (cf. Bamboo (2010): 7-8, and Unsworth (2000)).
Selecting	Adding objects of interest or information to the corpus or removing elements from the corpus based on certain criteria. Selecting modifies an existing corpus by removing and adding elements or constitutes a new corpus by adding the first element to it.
Writing	Proper writing, e.g. of a draft for a journal article or a thesis chapter (cf. Palmer et al. (2009): 21).
Assembling	Putting any kind of elements from the corpus together to form a work which can be shared, published or disseminated. An iterative and continuous process which is based on or may involve other Activities such as writing, reading, sampling etc. (cf. Palmer et al. (2009): 22).
Notetaking	Jotting down thoughts, remarks or notes at any stage of the working and research process and independently from particular objects of interest (cf. Palmer et al. (2009): 10-11).
Illustrating	Visualising an idea, an argument, a relationship or context expressed by text, speech or other visual aids (cf. Bamboo (2010): 8, and Unsworth (2000)).
Sharing	Making (intermediate) research results available to a (selected) audience such as a working group (cf. Bamboo (2010): 9-11).
Publishing	Making (intermediate) research results available to a wider audience such as the general public (cf. Bamboo (2010): 9-11).
Disseminating	Making (intermediate) research results available on a more collaborative, continuous and social basis such as attending and speaking at meetings, conferences, scholarly associations and societies (cf. Palmer et al. (2009): 23-25).

Table 1. Scholarly Activities.

2.7 References

- **Anderson, S. et al.** (2010). Methodological commons: arts and humanities e-Science fundamentals. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 368 (1925), pp. 3779–3796. http://rsta.royalsocietypublishing.org/content/368/1925/3779.short (accessed 31 January 2015).

- **Atkins, D. E. et al.** (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Michigan, MI: University of Michigan Library Press. http://www.nsf.gov/cise/sci/reports/atkins.pdf (accessed 31 January 2015).

- **Bamboo Project (2010)**. Project Bamboo Scholarly Practice Report. https://wikihub.berkeley.edu/download/attachments/68619618/bamboo_scholarly_practice_report.pdf?version=2&modificationDate=1292887646732 (accessed 31 January 2015).

- **Benardou, A. et al.** (2010). A Conceptual Model for Scholarly Research Activity. In Reilly, M. (ed), IConference Papers 2010. Champaign, IL: University of Illinois Press. https://www.ideals.illinois.edu/handle/2142/14879 (accessed 31 January 2015).

- **Benardou, A. et al.** (2013). Deliverable D1.2 – State of the art report on digital research practices, tools and scholarly content use. Europeana Cloud Deliverable. http://pro.europeana.eu/documents/1414567/a0a84066-a601-4737-945e-ab63484ae804 (accessed 31 January 2015).

- **Beynon, M.; Russ, S. and McCarty, W.** (2006). Human Computing: Modelling with Meaning, Literary and Linguistic Computing, 21 (2), pp. 141–157: 10.1093/llc/fql015 (accessed 31 January 2015).

- **Blanke, T. and Dunn, S.** (2006). The Arts and Humanities e-Science Initiative in the UK. In Sloot, P. M. A (ed), Second IEEE International Conference on e-Science and Grid Computing, Amsterdam, NL, December 2006. Los Alamitos, CA: IEEE Computer Society, p. 136: 10.1109/E-SCIENCE.2006.261069 (accessed 31 January 2015).

- **Blanke, T. and Hedges, M.** (2013). Scholarly Primitives Building institutional infrastructure for humanities e-Science. Future Generation Computer Systems, 29 (2), pp. 654–661: 10.1016/j.future.2011.06.006 (accessed 31 January 2015).

- **Borgman, C. L.** (2007). Scholarship in the Digital Age: Information, infrastructure, and the Internet. Cambridge, MA: Massachusetts Institute of Technology Press.

- **Brockman, W. S.** (2001). Scholarly work in the humanities and the evolving information environment. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub104/pub104.pdf (accessed 31 January 2015).

- **Bush, V.** (1945). As We May Think. Atlantic Magazine. http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ (accessed 31 January 2015).

- **Candela, L., Castelli, D. and Pagano, P.** (2013). Virtual Research Environments: An Overview and a Research Agenda, Data Science Journal, 12, [no pagination]: 10.2481/dsj.GRDI-013 (accessed 31 January 2015).

- **Deegan, M. and McCarty, W.** (eds) (2012). Collaborative Research in the Digital Humanities: A volume in Honour of Harold Short on the Occasion of his 65th Birthday and his Retirement, September 2010. Farnham: Ashgate.

- **Di Donato, F. et al.** (2013). Semantic annotation with Pundit: a case study and a practical demonstration. In Tomasi, F. and Vitali, F. (eds), Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment metadata, vocabularies and techniques in the Digital Humanities, New York, NY: Association for Computing Machinery, [no pagination]: 10.1145/2517978.2517995 (accessed 31 January 2015).

- **Doerr, M. et al.** (2011). Factual argumentation – a core model for assertions making. Journal on Computing and Cultural Heritage, 3 (3), pp. 1–34: 10.1145/1921614.1921615 (accessed 31 January 2015).

- **Dörk, M. et al.** (2012). Visualizing explicit and implicit relations of complex information spaces. Information Visualization, 11 (1), pp. 5–21. http://ivi.sagepub.com/content/11/1/5.full.pdf+html (accessed 31 January 2015).

- **Dombrowski, Q.** (2014). What Ever Happened to Project Bamboo?, Literary and Linguistic Computing, 29(3), pp. 326–339: 10.1093/llc/fqu026 (accessed 31 January 2015).

- **Drucker, J.** (2012). Humanistic Theory and Digital Scholarship. In Gold, M. K. (ed) Debates in the Digital Humanities, Minneapolis, MN: University of Minnesota Press, pp. 85–95. http://dhdebates.gc.cuny.edu/debates/text/34 (accessed 31 Janaury 2015).

- **Gibbs, F.** (2013). Digital Humanities Definitions by Type. In Terras, M., Nyhan, J., and Vanhoutte, E., (eds), Defining Digital Humanities. A Reader. Farnham: Ashgate, pp. 289-297.

- **Glock, H.-J.** (1996). A Wittgenstein dictionary. Oxford: Blackwell.

- **Gold, M. K.** (ed) (2012). Debates in the Digital Humanities, Minneapolis, MN: University of Minnesota Press.

- **Gradmann, S.** (2010). Knowledge = Information in Context : on the Importance of Semantic Contextualisation in Europeana. http://de.scribd.com/doc/32110457/Europeana-White-Paper-1 (accessed 31 January 2015).

- **Gradmann, S. and Meister, J. C.** (2008). Digital document and interpretation: re-thinking “text” and scholarship in electronic settings, Poiesis & Praxis, 5 (2), pp. 139–153: 10.1007/s10202-007-0042-y (accessed 31 January 2015).

- **Gradmann, S.** (2013): Linked Data Scholarship: Modeling and Interpretation in the Digital Humanities. Presentation held at Universidad Carlos III de Madrid, July 2013. http://www.slideshare.net/gradmans/20130711-linked-datascholarshipmadrid (accessed 31 January 2015).

- **Grassi, M. et al.** (2012). Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries. In Mitschick, A. et al. (eds), Semantic Digital Archives 2012: Proceedings of the Second International Workshop on Semantic Digital Archives, Paphos, CY, September 2012. http://ceur-ws.org/Vol-912/paper4.pdf (accessed 31 January 2015).

- **Grassi, M. et al.** (2013). Pundit: augmenting web contents with semantics. Literary and Linguistic Computing, 28 (4), pp. 640–659: 10.1093/llc/fqt060 (accessed 31 January 2015).

- **Harper, S. et al.** (eds) (2007). Proceedings of the eighteenth conference on Hypertext and hypermedia, New York, NY: Association for Computing Machinery.

- **Hughes, L. (2013).** NeDiMAH: Network of Digital Methods in the Arts and Humanities. Presentation held at Digital Humanities Luxembourg, December 2013. http://de.slideshare.net/lorna_hughes/lorna-hughes-12-052013-nedimah-and-ontology-for-dh (accessed 31 January 2015).

- **Mácha, J., Falch, R. J. and Pichler, A.** (2013). Overlapping and Competing Ontologies. In In Tomasi, F. and Vitali, F. (eds), Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment metadata, vocabularies and techniques in the Digital Humanities, New York, NY: Association for Computing Machinery, [no pagination]: 10.1145/2517978.2517984 (accessed 31 January 2015).

- **McCarty, W.** (2002). Humanities Computing: Essential Problems, Experimental Practice, Literary and Linguistic Computing, 17 (1), pp 103-125: 10.1093/llc/17.1.103 (accessed 31 January 2015).

- **McCarty, W. and Short, H.** (2002). Mapping the Field: Report of ALLC meeting held in Pisa, April 2002. http://www.allc.org/node/188 (accessed 31 January 2015).

- **McCarty, W.** (2004). Modeling: A Study in Words and Meanings. In Schreibman, S., Siemens, R. and Unsworth, J. (eds), A companion to digital humanities. Malden, MA: Blackwell. http://www.digitalhumanities.org/companion/ (accessed 31 January 2015).

- **McCarty, W.** (2005). Humanities computing. Houndmills: Palgrave Macmillan.

- **Meister, J. C.** (ed) (2012). Digital Humanities 2012: Conference Abstracts, University of Hamburg, July 16-22. Hamburg: University Press.

- **Mitschick. A. et al.** (eds) (2012). Proceedings of the Second International Workshop on Semantic Digital Archives, Paphos, CY, September 2012.

- **Morbidoni, C. et al.** (2011). Introducing the Semlib project: semantic web tools for digital libraries. In Predoiu, L. et al. (eds), Proceedings of the 1st International Workshop on Semantic Digital Archives, Berlin, DE, September 2011, pp. 97–108. http://ceur-ws.org/Vol-801/paper9.pdf (accessed 31 January 2015).

- **Morbidoni, C. et al.** (2013). Semantic Augmentation and Externalization in the Humanities: a Demonstrative Use Case. In The European Association for Digital Humanities et al. (eds), Digital Humanities 2013: Conference Abstracts, University of Nebraska-Lincoln, USA, July 2013. Lincoln, NE: Center for Digital Research in the Humanities, pp. 316-320. http://dh2013.unl.edu/abstracts/ab-337.html (accessed January 2015).

- **Moulin, C. et al.** (eds) (2011a). Research Infrastructures in the Digital Humanities: Executive Summary. www.esf.org/fileadmin/Public_documents/Publications/spb42_ExecSum.pdf (accessed 31 January 2015).

- **Moulin, C. et al.** (eds) (2011b) Research Infrastructures in the Digital Humanities. http://www.esf.org/fileadmin/Public_documents/Publications/spb42_RI_DigitalHumanities.pdf (accessed 31 January 2015).

- **Nucci, M. et al.** (2012). Enriching Digital Libraries Contents with SemLib Semantic Annotation System. In Meister, J. C. (ed), Digital Humanities 2012: Conference Abstracts, University of Hamburg, July 16-22. Hamburg: University Press, pp. 318–321. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/enriching-digital-libraries-contents-with-semlib-semantic-annotation-system.1.html (accessed 31 January 2015).

- **Oldman, D., Doerr, M. and Gradmann, S.** (n.d.). ZEN and the Art of Linked Data. New Strategies for a Semantic Web of Humanist Knowledge. To be published in Schreibman, S., Siemens, R. and Unsworth, J. (eds), A new Companion to Digital Humanities. Oxford: Blackwell [preprint].

- **Palmer, C. L., Teffeau, L. C. and Pirmann, C. M.** (2009). Scholarly information practices in the online environment: Themes from the literature and implications for library service development. Dublin, OH: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2009/2009-02.pdf (accessed 31 January 2015).

- **Pichler, A. and Zöllner-Weber, A.** (2012). Towards Wittgenstein on the Semantic Web. In Meister, J. C. (ed), Digital Humanities 2012: Conference Abstracts, University of Hamburg, July 16-22. Hamburg: University Press, pp. 318–321. http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/towards-wittgenstein-on-the-semantic-web.1.html (accessed 31 January 2015).

- **Poole, A. H.** (2013). Now is the Future Now? The Urgency of Digital Curation in the Digital Humanities. Digital Humanities Quarterly, 7(2). http://www.digitalhumanities.org/dhq/vol/7/2/index.html (accessed 31 January 2015).

- **Predoiu, L. et al.** (eds) (2011). Proceedings of the 1st International Workshop on Semantic Digital Archives, Berlin, September 2011.

- **Reilly, M.** (ed) (2010). IConference Papers 2010. Champaign, IL: University of Illinois Press.

- **Rockwell, G.** (2010). As Transparent as Infrastructure: On the research of cyberinfrastructure in the humanities. http://cnx.org/content/m34315/1.2/ (accessed 31 January 2015).

- **Schraefel, M. C.** (2007). What is an Analogue for the Semantic Web and Why is Having one important? In Harper, S. et al. (eds), Proceedings of the eighteenth conference on Hypertext and hypermedia, New York, NY: Association for Computing Machinery, pp. 123–132: 10.1145/1324960.1324966 (accessed 31 January 2015).

- **Schreibman, S., Siemens, R. and Unsworth, J.** (eds) (2004). A companion to digital humanities. Malden, MA: Blackwell. http://www.digitalhumanities.org/companion/ (accessed 31 January 2015).

- **Schreibman, S. et al.** (2013). Beyond Infrastructure: Modelling Scholarly Research and Collaboration. In The European Association for Digital Humanities et al. (eds), Digital Humanities 2013: Conference Abstracts, University of Nebraska-Lincoln, USA, July 2013. Lincoln, NE: Center for Digital Research in the Humanities, pp. 386-289. http://dh2013.unl.edu/abstracts/ab-276.html (accessed 31 January 2015).

- **Sloot, P. M. A.** (ed) (2006). Second IEEE International Conference on e-Science and Grid Computing, Amsterdam, NL December 2006. Los Alamitos, CA: IEEE Computer Society.

- **The European Association for Digital Humanities et al.** (2013). Digital Humanities 2013: Conference Abstracts, University of Nebraska-Lincoln, USA, July 2013. Lincoln, NE: Center for Digital Research in the Humanities.

- **Thaller, M.** (2013). Praising Imperfection: Why editions do not have to be finished. Lecture held at Culture & Technology - The European Summer School in Digital Humanities, Leipzig, July 2013. http://www.culingtec.uni-leipzig.de/ESU_C_T/node/292 (accessed 31 January 2015).

- **Terras, M., Nyhan, J., and Vanhoutte, E.** (eds) (2013). Defining Digital Humanities. A Reader. Farnham: Ashgate.

- **Tomasi, F. and Vitali, F.** (eds) (2013). Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment metadata, vocabularies and techniques in the Digital Humanities, New York, NY: Association for Computing Machinery.

- **Unsworth, J.** (2000). Scholarly Primitives: What methods do humanities researchers have in common, and how might our tools reflect this? http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html (accessed 31 January 2015).

- **Unsworth, J. et al.** (2006). Our Cultural Commonwealth: Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. New York: American Council of Learned Societies. http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf (accessed 31 January 2015).

- **Unsworth, J. and Tupman, C.** (2012). Interview with John Unsworth, April 2011, carried out and transcribed by Charlotte Tupman. In Deegan, M. and McCarty, W. (eds), Collaborative research in the digital humanities: A volume in honour of Harold Short on the occasion of his 65th birthday and his retirement. September 2010. Farnham: Ashgate, pp. 231–240.

- **Williams, D.-A.** (n.d.). Method as tautology in the digital Humanities, to be published in Digital Scholarship in the Humanities [preprint].

3 Report on Interviews

In order to collect empirical evidence for the validity of our Scholarly Domain Model, 16 semi-structured interviews were conducted during the end of 2013. In collaboration with the base project “Virtual and Real Architecture of Knowledge” (VRAK) within the Cluster of Excellence "Image, Knowledge, Gestaltung”[64] at Humboldt-Universität zu Berlin, a guideline for semi-structured interviews was devised. Interviews were meant to provide input for the creation of the use cases which would cover exemplary research and other scholarly processes. The method of the semi-structured interview is a conversation between two individuals which is structured by a list of questions. The list need not adhered to the letter, but both parties are free to digress and provide deeper aspects of the topics to be discussed. As quantitative methods cannot be applied in the evaluation of the results, it makes the analysis more difficult. However, it promises to yield more information between the lines. Both parties are free to digress and provide deeper aspects of the topics to be discussed and the interviewee has the opportunity to actually reflect on her own work freely.

Of the 16 interview partners, 10 were research scholars and 3 professors and remainder being an assistant professor, a student and a public relations manager with an average age of 36,6 years (median 33,5 years). In terms of subjects, the population was fairly evenly distributed over the humanities (6), the sciences (5) and engineering (5). In most cases (14), the mother language was German, two were native English speakers. The interviews were conducted in the mother tongue of the interview partners. In all cases but one, the interviews took place with both parties being present in one room, one interview was conducted remotely. Each interview lasted between 45 and 120 minutes, the audio was recorded for later evaluation.

Given the rather small population, the outcome of these interviews cannot be seen as representative of their respective discipline, but it nevertheless gives insight into the everyday practice of scholarly work and its surrounding activities, providing valuable input to the overall design of the SDM.

3.1 Comments on the questions asked

The guidelines (the English version can be found in the appendix in 3.4 Appendix: Interview Guideline) was modelled roughly after the classes and the general layout of the SDM in order to relate the methods and activities of the scholars directly to it.

After a short section of questions about the age and work of the interviewee, the first set of questions concerned the typical tasks the scholar faced in her everyday work. The aim here was to list up all the tasks of a normal working day and to rate it by factors like importance, urgency, recurrence, frequency and degree of favour.

Everyday work was also subject of the following part which was about the organisation and planning of work. The purpose of that was to find out among other things if there is a kind of understanding of a categorisation of tasks that might reflect the areas of the SDM or the activities therein.

Focussing more on the areas, we continued with asking questions about research and, more specifically, search strategies and after that excerption strategies which also covered the building of a corpus, the annotation of the sources and also the choice of software and specific methods in that respect.

[64] http://www.interdisciplinary-laboratory.hu-berlin.de/en

The next set of questions was about visualisation, contextualisation and reasoning. We wanted to know about the tools employed here and if such things are at all part of the scholars” research. After that, we asked for a summary of all the devices and software that are used at work and also asked for what kind of software is missing.

The last two sections were about collaboration and publication and respectively covered aspects of Social Context and Externalisation.

3.2 Findings

The results of the interviews are grouped here by the five principal areas of the Scholarly Domain Model with the remainder coming at the end.

3.2.1 Input

The question was how the research of a new topic was tackled. One interviewee stated that researching never starts at zero, because a new topic is always based on a previous problem. This underlines our concept that Output is connected with Input, albeit in another iteration of the cycle. Another scholar stressed the benefit of online databases and encyclopaedias which enables her to compile a basic bibliography (which we would consider as the corpus) within one day.

As for search strategies, the methods used here are quite diverse. While online tools are used for building bibliographies in favour of browsing through proper books, many interview partners mentioned that information from colleagues was valued higher ("more fruitful") than the search in databases, because it gives weighted and qualified opinions. This stresses of course the collaborative aspect in the Social Context. What follows after that initial search is the use of chaining by using references from the relevant articles. Search goes often together with the use of filters (e. g. genre) in the online databases. When asked if it would be helpful to be assisted by a semantic mark-up when searching (like searching for Albert Einstein but connecting that string also to the condition that he has the role of an author), one informant thought of it as being unnecessary as the intellectual effort for implementing it was deemed too high. Clearly it must be stated that technologies like faceted browsing already implement this and is also used successfully in search interfaces.

When asked about the choice of search terms, it became clear that the approach is in the whole fairly basic and that a lot of effort is still intellectual work: by the majority of the informers, basic search terms are preferred, combined with a manual leafing through results. Keyword search in databases was also mentioned. When asked for the reaction if there were no satisfactory hits, one interview partner answered "you always get something", meaning that there are always some traces that can be picked up from the results. Another approach is the refinement of the search, a broader term or the use of a different database.

In terms of the exploration of the Corpus, the use of available search capabilities was emphasised. Here the same search strategies apply as above, meaning that it is easier and faster to scroll through results of a dumb search (referred to as "human pattern spotting" by one informant) than to think of sophisticated search mechanisms like regular expressions. It was also felt by a few interview partners that the application of keywords was superfluous. "Tags multiply", commented a scholar on this, meaning that own tagging systems can become inconsistent quite easily. "Why duplicate the computer’s efforts?", was asked by another, referring to the ability of computers to quickly find data again.

In respect to the whole complex of searching, worries were mentioned by two informants about the fact that internet search engines are not necessarily designed to fit all needs of a scholar but also might be designed to help users find pages again that they already visited. The algorithm that produces the results is in general unknown to the users. Also, worries were mentioned about the predominance of English resources.

3.2.2 Research

The Research area was mainly covered by the question of how and where data is annotated. It was responded by the majority that annotation is an "ongoing part throughout the whole process", but one informant thought that the system of keeping annotations is still rather incoherent: Thoughts would be kept in notebooks or digitally, but not during reading itself, and most of the informants claimed that they would not write in books (although all of them stated that having access to books with annotations from famous scholars can be very revealing). Thus, the plethora of available tools leads to the situation that annotations end up all over the place (Zotero, textfiles, Evernote among others) with the possibility that annotations get lost or are invisible if e.g. a PDF file is opened with a different reader.

Annotations are also kept on paper (one scholar noted that she prints out important articles and also files them physically) and are then out of the digital realm. In other cases, extensive notes are made on printouts, e.g. when commenting on papers by students.

Another question asked was the one what Reasoning might mean to the interview partners. Asking the question to native speakers of English who have a background as historians or mathematicians resulted in rather generic answers (e.g. "working from some sort of evidence to reach a conclusion" or "that’s what we do, thinking through things"), while one German speaker (the English term "Reasoning" was used here, as well) related it to spatial reasoning, rationality and mental models, but also, more generally as “argumentation”, translating it as “Vernunft, Begründen”. Four informants had no clue.

3.2.3 Output

In terms of Externalisation, several output methods were discussed e.g. the inclusion of diagrams, graphs and functions in printed works. One interview partner mentioned that the visualisation of optical experiments would have been beneficial in a publication, but that it was impossible to implement interactive elements in paper publications. On the other hand, linking in online publications was mentioned very often.

It was also mentioned by a scholar working in the USA that sole electronic publishing is not recognised as valuable output.

3.2.4 Social Context

Hints to the importance of including the Social Context in a model of scholarly work have already shown up in the previous parts (e.g. asking colleagues for literature), but it was directly addressed by asking about the collaborative work they were engaged with.

In general, collaborative work is appreciated as long as the team is working well together and, emphasised by one scholar, there are phases of working together physically. Again, it was mentioned, that in some disciplines, collaborative efforts are discouraged and do not count as scholarly work. Disadvantages mentioned in terms of collaboration were the loss of control about the whole process and that that way of working is more time-consuming.

3.2.5 Documentation

In terms of Documentation, few interview partners stated that they kept a journal for themselves. Rather, documentation was primarily done for administrative purposes, annual reports of the institution or reports to funding agencies. One scholar noted that she kept a journal for just one single project, because of the size of the project and the fact that it employed new methods that were new to her. Also, it was referred to the fact that in that project a version control system was used which also contributes to the documentation.

3.2.6 Additional

Two aspects that are not directly covered by the SDM were mentioned quite often by the interviewees: teaching and administrative work. As was made clear in the first part of the interviews, both can take over a huge part of the ordinary working day. Clearly these are tasks that are to be distinguished from scholarly work, but they do influence it nevertheless. Furthermore, they might be located in the fields of Social Context and Documentation. Teaching additionally carries aspects of Externalisation.

We conclude that, although not being unrelated to scholarly work, the SDM will not be extended to cater for the special needs that arise through teaching and administration. The SDM will remain as a means for modelling research and consists of research primitives. We assume that teaching takes place predominantly in the Social Context and administration takes place in Documentation. As can be seen in the schema of the SDM, the five areas overlap, and it must be also understood that the importance of specific areas can change, depending on the current usage of the model.

3.3 Conclusion

The purpose of conducting these interviews was to add a bottom-up view of how research is actually performed by scholars. Clearly, the amount of scholars asked was a rather small sample, but it nevertheless yielded a few substantial underpinnings for the construction of our model.

The first one is the fact that the social context in which research is performed plays an important role. This is certainly the case when asking colleagues for literature and other resources rather than searching on one’s own, but also when collaborative projects are at hand. Here, of course, digital tools play an important role, because they clearly have facilitated sharing of material, collective production of text and accelerated communication. Also, although there are caveats, collaborative work is seen as fruitful.

Secondly, it became clear that exploration of new topics is still a matter of intellectual work. The informants agreed on the fact that a search with basic terms is in most cases sufficient and the manual leafing through a longer list of results is quicker than thinking of complex searches. Moreover, a broader search might also reveal chance findings that might have slipped through the net. This is another justification for us to separate the Exploration from Searching, as the Search for something that is already known but has to be found again is completely different from that.

Thirdly, the use of annotation is of course a technique that is an integral part of research. With regard to its implementation in the digital domain, there is still work to be done. At the moment, too many systems exist which makes it difficult to keep track of where annotations are stored and also who is able to see them, when stored online.

As a final observation from our interviews, it became clear that time for proper research is delimited by other obligations in scholarly life, namely administrative tasks and teaching. Even though the amount of time dedicated to administrative tasks might not take up much of the time, it can still be perceived as an intrusion and a distraction. One exception being the social aspect of meetings. Teaching, on the other hand, was mentioned as being quite enjoyable, and this surely has to do with the aspect of being in a social context and at the same time being able to externalise elements of one's own research.

3.4 Appendix: Interview Guideline

About the interviewee

- What year were you born?
- In what field or discipline are you trained?
- Where do you work?
- What is your academic status?
- What are your main research interests?
- What would you say is your most important professional function?
- When did you start working with computers?

Types of tasks, typical everyday tasks (as table)

- Kind of task (How would you call that task?)
- Importance (How important is this task compared to other tasks?)
- Recurrence (Is this a repeating task, if yes in what interval?)
- Estimation (How much time do you estimate for this task?)
- Urgency (How urgent is completing this task typically?)
- Frequency (How often do you work on this task?)
- Influenced from outside (Do you work on this task on your own account or is it controlled from the outside (Deadline/Meeting)?
- Duration without break (How long do you work on this task typically without a break?)
- Task is combined with… {Kind of task} (Which tasks is this task typically combined with (e.g. reading and excerpting, meeting and taking notes/protocol)
- Like this task (How much do you like to do this task?) 1-5 where 5 is best.

Organization / Planning

- How do you plan tasks and dates?
- How do you prioritize tasks? (Importance/Urgency)
- How do you document your work?

- How often do you update your plans or documentation of work?
- How realistic is your scheduling?
- How do you organize your notes and information regarding to your workspace? (files on the desktop, shelves, folders, papers with notes)?

Research

- **Typical research questions**
  - What overall questions are the integral part of your work?
  - How do you get to your specific research questions?
  - Are there recurring research questions?
- **Research strategies/ research sources**
  - When doing research on a topic, which resources do you use frequently (both analogue and digital), (historical dictionaries, special dictionaries, online dictionaries)?
  - How do you proceed while researching?
  - What exactly are you looking for?
  - What do you do if you fail to find something?
  - Do you work interdisciplinary?
  - What methods do you use while working?

Excerption strategies

- What genres of text do you read on screen and which do you print?
- How do you organize ideas, thoughts and other relevant information?
- How do you excerpt from texts?
- How do you manage your collection of texts?
- How do you organize your excerpts?
- What do you want to know about a text apart from its contents? Where would you store it? (Important metadata)

Visualization, Contextualization and Reasoning

- Do you use any types of visualization and for what purpose?
- Is it important for you to visualize connections? And is that helpful?
- Do you know and or use visualizing tools?
- Do you use visually refined data connections in your work and publications?
- What types of visualization do you especially like?

- When visualizing objects, what kind of contextual information would you like to see there?
- What does the term "reasoning" mean to you?

Tools

- What electronic devices do you use? (smartphone, laptop, tablet, desktop computer, ebook-reader)
- Which programs are most important for your work?
- What other programs do you use frequently apart from that?
- Which tools do you like and not like?
- Do you use any systems of classification or ontologies?
- Can you think of tools (fictitious or real) that would facilitate or improve your work?

Collaboration and Communication (social context)

- When you collaborate with other people at work, how do you manage to do that?
- How often do you collaborate (e.g. in projects, publications, talks)?
- Do you like working collaboratively?

Publishing (output)

- Do you publish more in print or more digitally?
- What kind of publication do you prefer?
- What does the term "Digital Humanities" mean to you?
- Would you consider yourself a scholar of digital humanities? Why?

4 Development of Pundit

In this section we report on the latest developments of the DM2E scholarly research platform, based on Pundit, Ask, Feed and Korbo. The first versions of such components were first documented in D3.2 Prototyping Platform Implemented. In D3.3 E-Learning Courses published, a user tutorial and documentation has been collected (and published in the project wiki at http://wiki.dm2e.eu/). In this section we focus on the main improvements over the already documented versions, which were informed by the feedback from the Digital Humanities Advisory Board (DHAB) and scholars participating in the experiments reported in this deliverable.

Even if the DM2E project is reaching its end, the development of the components is still ongoing, as the tools are used in other Digital Humanities related projects as well as constitute an important business asset at Net7.

4.1 Pundit

4.1.1 New web based user interface: Pundit 2

Web UIs quickly become obsolete. In order to make the tool more attractive and stable the client side, the Pundit UI was re-built using AngularJS[65], a cutting edge JavaScript framework from Google. The new UI is the result of a usability and user interaction study performed by Net7 in collaboration with the Cluster of Excellence at Humboldt-Universität zu Berlin, "Image, Knowledge, Gestaltung".[66] The result is two-fold. On the one hand it makes the UI more attractive and usable, on the other hand it makes it more extensible and more manageable in terms of changes and adaptations.

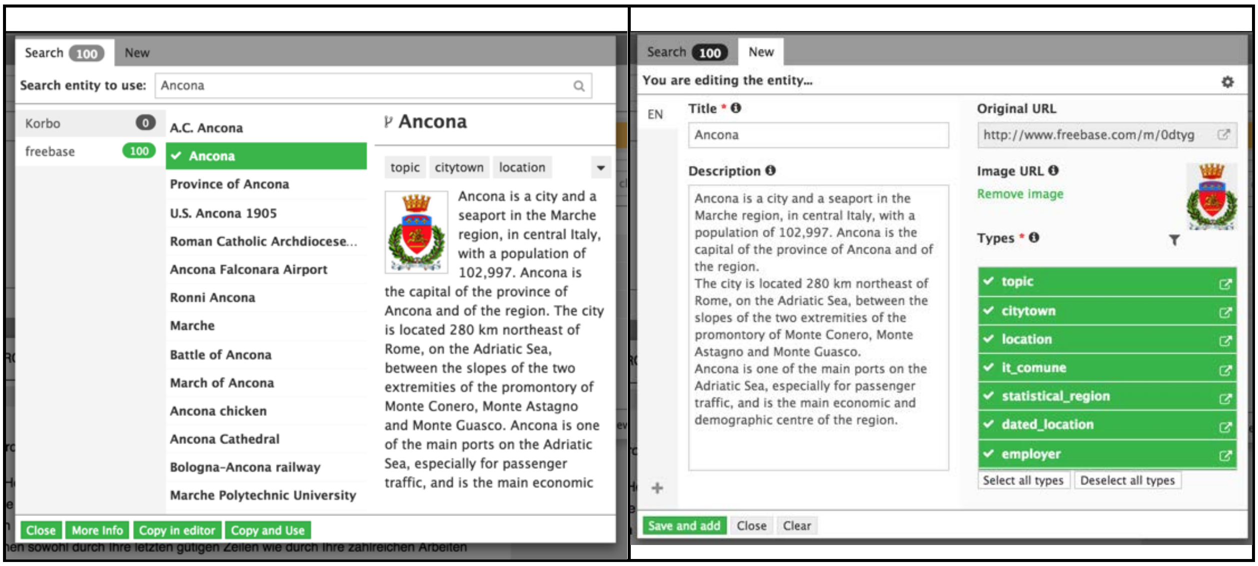
An important outcome of this activity was that of integrating into the same UI a number of features that were previously distributed across different web applications (e.g. Korbo and Ask). This responds to a practical need, expressed several times by scholars using the platform and by the DHAB. Scholars find themselves more comfortable using a single UI interface instead of being forced to open different UIs to perform tasks that are perceived as phases of the same conceptual workflow.

The main improvements are:

- **Editing annotation vocabularies on-the-fly.** The ability to create and easily maintain annotation vocabularies is important to iteratively model Scholarly Operations in a given domain (as pointed out by experiments results). To make this easier, editing annotation vocabularies (e.g. adding and modifying entries) is now possible from within the Pundit 2 UI, without switching to a separate system (e.g. Korbo), as shown in figure 6. This is done transparently by interacting with the Korbo server REST API. However, while the UI supports easy addition of instances of predefined classes, in the current version, adding classes and properties can only be done by editing a JSON configuration file. Improving the UI in this direction is a requirement that will be taken into account for the next development phase of the tool.

[65] https://angularjs.org/

[66] http://www.interdisciplinary-laboratory.hu-berlin.de/en



Ancona is a city and a seaport in the Marche region, in central Italy, with a population of 102,997. Ancona is the capital of the province of Ancona and of the region. The city is located 280 km northeast of Rome, on the Adriatic Sea, between the slopes of the two extremities of the promontory of Monte Conero, Monte Astagno and Monte Guasco. Ancona is one of the main ports on the Adriatic Sea, especially for passenger traffic, and is the main economic

Ancona is a city and a seaport in the Marche region, in central Italy, with a population of 102,997. Ancona is the capital of the province of Ancona and of the region.
The city is located 280 km northeast of Rome, on the Adriatic Sea, between the slopes of the two extremities of the promontory of Monte Conero, Monte Astagno and Monte Guasco.
Ancona is one of the main ports on the Adriatic Sea, especially for passenger traffic, and is the main economic and demographic centre of the region.

Figure 6. The new Pundit UI widget to add a new item into the annotation vocabulary.[67]

- **Managing personal notebooks.** Managing and sharing different personal notebooks is deemed as important with respect to scholarly areas such as Documentation and Social Context. Notebook management, including creating, deleting notebooks and switching from one notebook to another is now directly handled by the Pundit 2 UI, without the need to switch to a different systems (e.g. Ask). However, a direct link to Ask, where public notebooks from other users can be aggregated and explored is provided.

- **Improvement of the Image annotation facility.** Feedback from digital scholars highlighted limitations with respect to the image annotation facility in the previous version of Pundit. To overcome this, the image annotation UI module has been improved, as shown in figure 7. In particular:

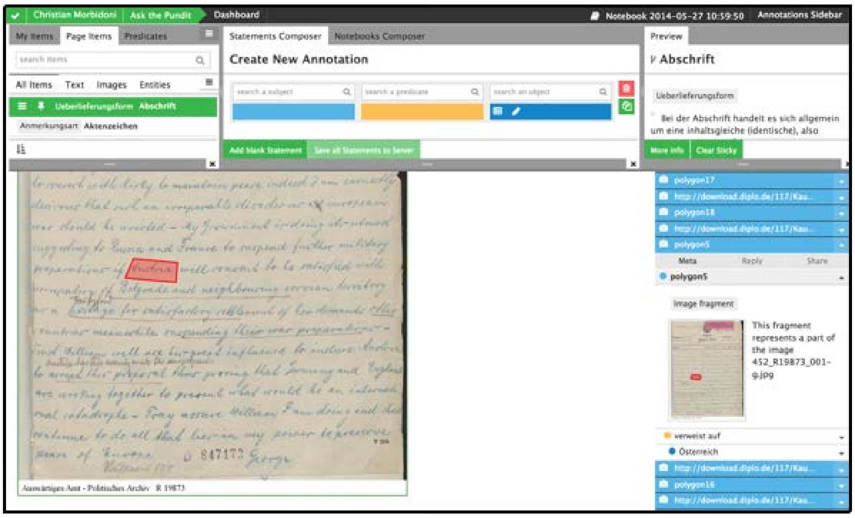
  - During the annotation creation process, full-screen mode is now supported to make it easier to annotate big images.

  - Zooming functionality has been improved.

  - Multiple polygons can be drawn on an image and the selections annotated.

  - Once an annotation on an image fragment is created, clicking on the annotation makes the annotated region (polygon) visible as an overlay on top of the image.

[67] New items can be first searched on a preconfigured LOD dataset (e.g. Freebase or DBpedia), copied into the annotation vocabularies and edited (e.g. adapting the description, choosing appropriate types/categories and depictions), or created from scratch.



Bei der Abschrift handelt es sich allgemein um eine inhaltsgleiche (identische), also

This fragment represents a part of the image 452_R19873_001-9.jpg

Figure 7. Screenshot of Pundit 2 showing an annotation on a fragment of image.

4.1.2 Configurability

The aim of the project was not to produce a single monolithic tool, but rather a tool that can be configured and deployed to be useful in different contexts and in different scholarly communities. That is why a lot of effort has been spent on improving the configurability of Pundit.

A complete list of possible configurations of the tool can be found here: http://dev.thepund.it/download/client/last-beta/docs/#!/api/punditConfig

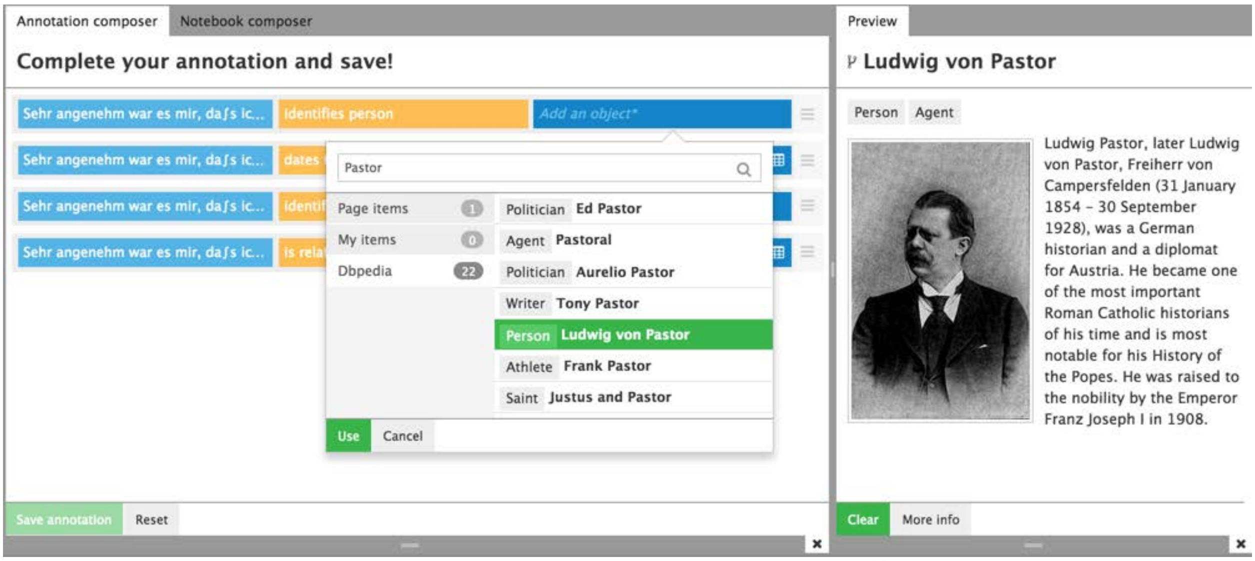
The ability to configure annotation vocabularies and relations as well as enabling/disabling specific UI modules was a key feature to enable the experiments documented in this deliverable. Making it possible for non-developers to quickly obtain and put online customised versions of the software (to then be used in specific experiments).

The configuration facility also allows to create new annotation templates (see next section) and make them immediately available to users.

4.1.3 New features

Beside improving the user interface and stabilising existing features, Pundit 2 also adds some new features. The main novelties are:

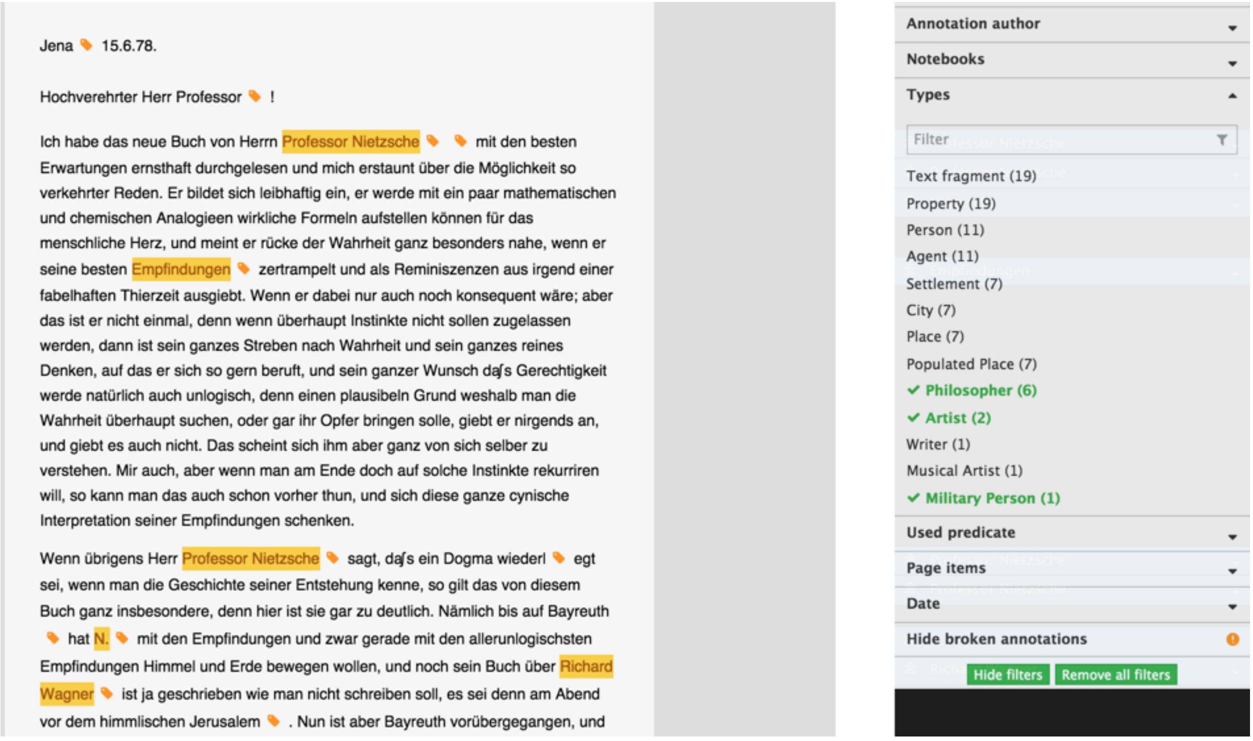
- **“Templating” annotations.** RDF based templates allows to model pre-defined types of annotations so that scholars can create them quickly. A template represents a specific annotation pattern and is defined by a set of RDF triples, where the objects, predicates and objects can be constrained to a specific value or left as non-grounded variables. When a template is selected, as soon as the user selects a text to be annotated, the annotations is automatically fed into the Pundit triple composer, where the user can use the search functions to select entities (e.g. from DBPedia or from a custom annotation vocabulary) to be associated to the non-grounded variables in the template. The ability to create new templates via a simple configuration and to make them immediately available to users is deemed important to create and iteratively model Scholarly Operations in a given domain (cf. report on SDM). Figure 8 shows the new template functionality.



Ludwig Pastor, later Ludwig von Pastor, Freiherr von Campersfelden (31 January 1854 – 30 September 1928), was a German historian and a diplomat for Austria. He became one of the most important Roman Catholic historians of his time and is most notable for his History of the Popes. He was raised to the nobility by the Emperor Franz Joseph I in 1908.

Figure 8. Annotation via templates in Pundit 2.[68]

- **Filtering annotations in a page.** Annotations in a page can now be filtered by author, date and other metadata. Annotations can also be filtered with respect to the type of entity involved in the annotation (see Figure 9).



Jena 15.6.78.

Hochverehrter Herr Professor !

Ich habe das neue Buch von Herrn Professor Nietzsche mit den besten Erwartungen ernsthaft durchgelesen und mich erstaunt über die Möglichkeit so verkehrter Reden. Er bildet sich leibhaftig ein, er werde mit ein paar mathematischen und chemischen Analogieen wirkliche Formeln aufstellen können für das menschliche Herz, und meint er rücke der Wahrheit ganz besonders nahe, wenn er seine besten Empfindungen zertrampelt und als Reminiszenzen aus irgend einer fabelhaften Thierzeit ausgiebt. Wenn er dabei nur auch noch konsequent wäre; aber das ist er nicht einmal, denn wenn überhaupt Instinkte nicht sollen zugelassen werden, dann ist sein ganzes Streben nach Wahrheit und sein ganzes reines Denken, auf das er sich so gern beruft, und sein ganzer Wunsch daſs Gerechtigkeit werde natürlich auch unlogisch, denn einen plausibeln Grund weshalb man die Wahrheit überhaupt suchen, oder gar ihr Opfer bringen solle, giebt er nirgends an, und giebt es auch nicht. Das scheint sich ihm aber ganz von sich selber zu verstehen. Mir auch, aber wenn man am Ende doch auf solche Instinkte rekurriren will, so kann man das auch schon vorher thun, und sich diese ganze cynische Interpretation seiner Empfindungen schenken.

Wenn übrigens Herr Professor Nietzsche sagt, daſs ein Dogma wiederl egt sei, wenn man die Geschichte seiner Entstehung kenne, so gilt das von diesem Buch ganz insbesondere, denn hier ist sie gar zu deutlich. Nämlich bis auf Bayreuth hat N. mit den Empfindungen und zwar gerade mit den allerunlogischsten Empfindungen Himmel und Erde bewegen wollen, und noch sein Buch über Richard Wagner ist ja geschrieben wie man nicht schreiben soll, es sei denn am Abend vor dem himmlischen Jerusalem . Nun ist aber Bayreuth vorübergegangen, und

Figure 9. Filtering annotations in a page in Pundit 2.

- **Anno-matic.** A number of named entity recognition services appeared in the last years and, despite the fact that their performances and precision is increasing, results

[68] Triples are automatically created from the template definition and non-grounded variables (e.g. the identified person in a sentence) can be quickly appointed by searching in available vocabularies and data sources.

are not always correct and automatic annotation of entity cannot be trusted without a manual revision from a domain expert. Pundit 2 includes support for automatic annotation by:

- Leveraging existing ER and linking web services. Built-in support is provided for Data TXT[69], supporting English and Italian languages, but other services can be supported by developing add-ons.
- Providing an easy to use UI for extracting entities from a selected text in a page and revising results (approving, rejecting entities matches) to create meaningful annotations.

4.1.4 Online resources

The official Pundit website is http://thepund.it. The source code of the Pundit client is available at https://github.com/net7/pundit2. The source code of the Pundit server side component is available at https://github.com/net7/pundit-server.

4.2 Other core components: Ask, Korbo, Feed

4.2.1 Korbo

Korbo is a server side component that provides APIs for handling collections and taxonomies of entities, and is used in connection with Pundit to store and handle annotation vocabularies.

While the previous version of Korbo included some UI modules (e.g. to edit vocabularies), the current version is designed as a pure REST API, and the main UI functionalities have been transferred in Pundit 2.

At the time of writing Korbo is being incorporated into the Pundit server side component that will constitute a single framework to store and manage annotations, vocabularies and entities collections.

The current version of Korbo is available at https://github.com/net7/korbo2.

4.2.2 Feed

Feed is a HTTP API component that exposes Pundit annotation environment as-a-service and is used make DM2E content annotatable with Pundit.

No significant changes were made to Feed with respect to the version documented in D3.3 (Consuming DM2E data in Feed).

The source code is available at https://github.com/net7/feed.

4.2.3 Ask

Ask is a web application allowing users to explore public notebooks and manage personal private ones. No significant improvements was made to Ask with respect to what

[69] https://dandelion.eu/products/datatxt/

documented in D3.3, but some of its features has been incorporated in Pundit 2 (as previously mentioned).

The source code is available at https://github.com/simonefonda/ask-pundit.

5 Report on Experiments

As part of Task 3.4, three experiments have been conducted with the semantic annotation application Pundit and additional components such as Korbo for simple vocabulary management and instance data creation and faceted browsers in the second half of 2014. While the experiments conducted in the context of the Wittgenstein Incubator focused on the usability of Pundit (cf. D1.2 – Final Integration Report), the experiments reported on here were designed to provide empirical input and add a practical bottom-up perspective to the more theoretical and top-down research regarding the functional primitives and Scholarly Operations as well as the “reasoning” (cf. introduction to the Deliverable).

The research interest of the experiments was to investigate how interpretative approaches of humanists can be operationalised in the particular context of Linked Data and Pundit and its components. For this purpose, humanists were confronted in real-life working contexts with the formal and explicit approach of Linked Data and semantic annotation.

The experiments particularly focused on the Scholarly Activities *annotating* and *visualising* both of which are seen as being pivotal to most humanists research activities. The aim was to investigate how these two activities materialise in different real-life use cases (cf. Scholarly Operations) focusing on interpretative approaches of humanists which have no prior knowledge of Linked Data or semantic annotation tools such as Pundit.

The principal topical and temporal horizon for the experiments was Historical Sciences and Contemporary History (19th/20th century).[70] Based on this precondition, during the first half of 2014, more than 70 historians and teachers mostly at German history departments and similar institutions were contacted and asked whether they personally or in the context of a seminar with students were willing to participate in the experiment. About 30 responses were received of which about 20 were positive and interested. From there we began investigating 8 different use cases and different topical orientations. After an initial round of deliberation with potential participants we chose 3 use cases for implementation.[71]

The three distinct use cases chosen for the experiment belong to the historical-archival domain. The first use case which has been created in cooperation with the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) and the Fachhochschule Potsdam (FHP) focused on the editorial and archival sciences. The use case created with the Georg-Eckert-Institut (GEI) focused on educational history. The use case devised with the historical seminar of the Humboldt-Universität zu Berlin (UBER) focused on visual history and the didactics of history, i.e. using tools such as Pundit for teaching history to students.

All three experiments and their respective use cases targeted the same overall research questions and were conducted by employing the same principal methodological approach. This principle plan for the experiments had been approved by the DHAB in its 4th meeting 3 April 2014.

The experiments investigated the following principal research questions:

- How can genuine research questions and interests be operationalised within the context of Linked Data and the particular context of Pundit?
    - o How do non-experts deal with Linked Data concepts and approaches?

[70] The main reason for choosing this topical and temporal orientation was that the two organisers of the experiments are trained historians which was expected to facilitate and support the experiments.
[71] The decision was based on different aspects such as the availability of appropriate digital material, the possibility to organise a seminar with students, and available time of the teachers and historians to invest into the experiments.

- How are the Scholarly Activities annotation, visualisation, and interpretative modelling reflected as Scholarly Operations?
    - Which “statements” do humanists consider necessary in their particular use case?
- Which potential do they see in Linked Data and Pundit for applications in the humanities?

The **methodological approach** and set-up of the experiments consisted of the following three principal stages:

1. Series of preparatory meetings with teachers/lecturers

   The first step in the preparation of the experiment was the identification of research questions or interests which were relevant to the given context of the use case. Then, the identified research questions and interests were operationalised for the context of Pundit and Linked Data in the form of simple annotation vocabularies. These meetings were part of the experiment and included open discussions and hand-written protocols as means of data recording.

2. Series of workshops and work at home

   The second phase of the experiments consisted of series of meetings with the students in the form of several workshops. In general, the participants were introduced to the particular seminar topics by their teachers. Afterwards, they were introduced to Linked Data and Pundit and the prepared annotation vocabularies were discussed. The students then worked with Pundit and the annotation vocabulary in the workshops as well as at home. During these workshops, data was collected by observational and interrogative means and recorded in hand-written protocols. At the end of each use case an extensive questionnaire was filled out by the participants.

3. Follow-up meetings with teachers

   The follow-up meetings were meant to reflect on the previous workshops and in particular to provide input on the future application scenarios and general advantages and disadvantages of Linked Data and Pundit. These meetings consisted of semi-structured discussions and were recorded by hand-written protocols.

In sum, data has been collected by open **interviews**, i.e. open discussions with the participants during every phase of the experiments, **observation** during the workshops including the research data created and collected in the notebooks, and via a common **questionnaire** at the end of each experiment. This questionnaire contained overall questions pertaining to the potential, shortcomings, advantages and disadvantages of Linked Data and Pundit as well as dedicated sections for each experiment.

All three use cases followed the principle set-up outlined above, however, each one also focused on additional and distinct aspects. The section “Use Cases” will introduce the topic and conduct of each use case in a broader perspective. Input for this section is based on the interviews and observations. The section “Questionnaire” will then take a more comprehensive perspective on the experiments, Linked Data, and Pundit by reporting on the most important results from the questionnaire. The section “Conclusion” will summarise the most relevant results and provide recommendations for engaging humanists with Linked Data in a fruitful and productive way.

5.1 Use Cases

In the following, all three use cases will be presented. The particular topics will be introduced and the preparatory phase and the conduct of the experiments discussed. In particular, we will comment on the creation and application of the annotation vocabularies.

For the experiments we had to operationalise each use case in the context of the current capabilities and state of Pundit and its components. With regard to the vocabulary creation that means we refrained from creating hierarchical class and property definitions since Pundit currently does not display these hierarchies. Therefore, we created flat annotation vocabularies. Also, we did not reuse classes or properties from existing ontologies for efficiency and time reasons. With regard to visualisation we also decided due to time constraints to utilise, and at the same time test, the available generic visualisation through the in-built faceted browser in Ask as well as simple custom-build faceted browser similar to the Wittgenstein Faceted Browser.[72]

5.1.1 Fachhochschule Potsdam (FHP)

The first use case stems from the discipline of Editorial and Archival Sciences. The workshop was held in collaboration with Markus Schnöpf, from the Berlin-Brandenburgische Akademie der Wissenschaften, who is involved in the Digital Humanities, in particular Digital History and Editorial and Archival Sciences and is associated with a lectureship for a seminar at the Fachhochschule Potsdam (FHP). All of the participants of the workshop, 15 in number, were Bachelor students attending the seminar “Editionstechniken” and the workshop was intended to complement the seminar with respect to digital techniques and methodology, in particular the application of Linked Data and the Pundit environment for Editorial and Archival Sciences. The Berlin School of Library and Information Science of the Humboldt-Universität zu Berlin hosted the workshop on two dates, 22 August and 12 September 2014.

The problem statement was if the creation of editorial guidelines for archival material are possible in a Linked Data context and if such guidelines in the form of a simple annotation vocabulary can be successfully applied in the context of Pundit. The topical focus of the use case was the crisis of July 1914. Diplomatic primary sources from the political archive of the German foreign ministry[73] were selected in advance by the teacher.

During the preparation of the workshop, a basic editorial guideline was devised in cooperation between the teacher and the organisers of the experiments. The guideline specified to mark-up (1) all textual phenomena concerning the structure of the documents, such as title, signatures, or remarks, (2) different scripts (“Hände”), (3) basic metadata such as author, date, or provenance, (4) persons and (5) places mentioned in the text, and (6) events referred to. Additionally, if possible, the participants were asked to relate the entities to existing entities from authority files like Virtual International Authority File (VIAF),[74] Geonames,[75] and DBpedia.[76] In the case of Geonames and VIAF, this was done manually by replacing URIs of instances in Korbo.

Based on this guideline, a simple core annotation vocabulary was prepared for the workshop. During the workshop, a group of students was asked to extend the vocabulary based on the guideline. One reason was that the conceptual creation of an editorial guideline, in this case the validation and possible extension of the prepared guideline, was

[72] Cf. http://metasound.dibet.univpm.it/dm2e/ajax-solr-master/examples/wab/
[73] http://www.archiv.diplo.de/
[74] http://viaf.org/
[75] http://www.geonames.org/
[76] http://dbpedia.org/

part of the students” seminar task. The other reason was that the particular focus of this use case was on the creation of a suitable vocabulary by the students themselves: are students able to translate the conceptual framework of editorial and archival sciences to a suitable vocabulary?

The vocabulary group extended the core vocabulary with text phenomena found in selected primary sources and based on the input by the other participants of the workshop. Considering the short timespan (1.5 weeks) for the creation of the extension, the specification of the vocabulary worked well: The most difficult conceptual problem for the participants was the differentiation between class and instance. A minor issue was the level of abstraction of the properties: several properties could have been subsumed under more general ones. Another difficulty for the participants was to determine whether they had created a comprehensive set of entities for the description of relevant phenomena in the primary sources.

After the group finished their specification, the lecturers implemented the vocabulary in Pundit. The participants then each selected one primary source and started working with the vocabulary based on the aforementioned guidelines. They were asked to document their work as well. Part of the working instructions were that the students had to not only provide a label for new instances but also a scope note describing the meaning of the new instance.

The appropriateness of the vocabulary has been proved by the fact that the other participants were able to apply the vocabulary during the workshops and that they did not ask for any significant additions during their work and also not in the questionnaire. Lastly, the actual triples created in the notebooks show that the participants did in fact successfully apply the editorial guidelines to the primary sources.

The translation of a simple editorial guideline to a RDF vocabulary proved to be possible during the preparatory phase and during the revision and extension phase during the workshop.

All the necessary or essential statements, as identified in the context of this seminar, for editorial work are factual statements, and, consequently, were easily representable in the triple structure. Examples of such factual statements, in contrast to more interpretative statements, are statements about structural text phenomena such as pages, signatures, titles etc., and, factual statements about the contents such as the author, topic, addressee etc.

A principal conceptual issue, which is not specific to the use case and which did not pose any practical problems during the workshop, are statements about the exact provenance of a digitised text: Is the provenance of the digital text the same as the analogue one? Another principal issue is the exact semantics of statements about the phenomena in the text: should a statement about the author of a text have the complete document and/or the proper name in the text as its subject? What exactly are we talking about when we refer to phenomena in the text, which are represented in the digital copy of this text?

The translation process also showed the chance to avoid overspecification (“über-diplomatisch”), i.e. focus on a basic core vocabulary for editing the sources, while still retaining the potential to extend and specify the vocabulary as needed. Whether editorial scientists should mark-up more or less phenomena in texts and whether users of edited texts profit from overly detailed mark-up remains open. The technical requirements towards the formalisations and implementation of an annotation vocabulary in Pundit, however, constitutes an opportunity to rethink these issues especially regarding open digital and networked working environments.

5.1.2 Georg-Eckert-Institut (GEI)

The second use case stems from the discipline of Educational History. In this case, only one participant, a trained historian from the Georg-Eckert-Institut (GEI) in Braunschweig, took part in the experiment. Furthermore, the experiment was not organised as a series of two workshops but the annotation work with Pundit was conducted over the course of several weeks in August and September 2014. Preparatory meetings and a follow-up meeting were held as in the case of the other two experiments.

The overall topic of the use case came from the project "World of Children"[77]. The research question was how children have been influenced and educated in their formative years in school. Investigating the formative years of adults yields important insights into how they think and write on the discourse on modernity. Textbooks are semi-official documents that were read by wider parts of the Germans during their formative years. With this material we try to find the representations of the world and the nation and the description of historical processes that were offered by the state to its future citizens. So, we search for representations of the nation and the globalised world. Also, we look for representations of change, crisis, religious conflict, social change and similar events.

The goal is to identify various topoi and their connotation and presentation in different kinds of school books: Which topoi appear in the different kinds of school books? How are they connotated and in which context are they put? These topoi will be compared over time, i.e. around 1850 and around 1900 in order to assess which and how specific topoi and their connotation change and which new ones appear or old ones disappear. For example, "nation", "globalisation", or "forming of the nation" are topoi which are discussed very differently in protestant and catholic school books. The connotation connected with topoi also differ: For example, "the Kaiserreich" is associated with backwardness and preventing the founding of a German nation. Lastly, the question is if these topoi and connotations can be grouped into specific "images of others" ("Fremdbilder") and "images of oneself" ("Selbstbilder")? The study is a qualitative analysis on the small scale. It may serve as a framework for subsequent and more extensive analysis (re-usability).

For the purpose of this experiment, the participant chose school books from the Digital Library of the GEI which are also available in the DM2E repository[78]. Annotations were made on the digitised pages of the chosen books.

This experiment additionally focused specifically on the aspect of reasoning. In contrast to the other use cases, the participant was explicitly asked to use ASK[79], a faceted browser for exploring annotations in notebooks created with Pundit, to try to explore new hypotheses based on filtering annotations. The results of this part of the experiment have been reported in the section on reasoning (cf. "Report on Reasoning").

During several meetings the research question and approach was developed and a basic annotation vocabulary of properties created. The participant then worked independently with Pundit over the course of several weeks in August and September 2014.

Since this particular use case had only one participant working over a longer time period, it was possible to create a detailed description of the actual working process. The first steps formally belong to the source critique in the historical methodology and included, in this particular case, the semi-random reading of the source material. Reference points for in-depth-view of the material were subheadings. Interesting text sections were annotated and

[77] http://www.dipf.de/de/dipf-aktuell/pdf-aktuelles/presseinformationen/pm-2014/PM_2014_29_04ProjektstartWeltderKinder.pdf
[78] http://data.dm2e.eu/data/html/dataset/gei/gei-digital/20140830013040893.
[79] http://demo-cloud.ask.thepund.it/

annotated fragments labelled. Then, triples were created about the source material regarding factual statements about the author, publication date, title, etc. After that, triples of second order were created which identified important historical persons and events. These subjects were then combined with either places, dates or states. Lastly, references were created to material outside of the corpus such as DBpedia in order to explain parts of the material to non-experts.

The focus of the vocabulary devised for the experiment is on the properties, i.e. the expressivity of the vocabulary stems from the different types of relations between the phenomena within the school books. In contrast to the first use case the phenomena marked up in the text in this use case are subject to interpretation in an extent that is significantly different from each other. The marked up phenomena consist mainly of connotations and subtle undertones. Irrespective of the phenomena as entities themselves, this enables the correlation of relations between them in a substantially more flexible way.

Yet, these phenomena can be expressed with the simple triple structure: Since the focus of these phenomena is on the type of relation between two entities, expressive properties can be utilised to express these relations.

The properties carry specific interpretations themselves merging several distinct statements into one property. For example, the property “is positively modern connotated with” is a complex statement embodying a hypothesis about the expected semantics of a text, i.e. between two entities. The semantic of the property expresses that in the context of a particular text an entity, the subject of triple, is presented as modern in a positive way by being discussed in the context of another entity, the object, which stands for modernity in a positive sense. For example, the “Reichsgründung” (founding of the Reich) “is positively modern connotated with” “Wirtschaftseinheit” (unity of economy). Both, the object and the subject are implicitly topoi.[80] This is a way to reduce complex semantics of interpretation to a reasonable level of abstraction.

In addition to that, interpretation is a potentially unending process of recurring succession, i.e. the certainty of particular statements made at one point during the working process may increase or decrease when confronted with new knowledge. This does not only apply to the material that is subject to interpretation but to the vocabulary as a means of interpretation itself. Thus it may be necessary to introduce new distinctions to annotate, beyond the annotation of what is immanent to the text initially worked with.

During the follow-up meeting, the participant stressed this fact again that there are layers or different levels of interpretation which may build upon each other. These levels should be clearly distinguishable in a network of statements, for example, statements immediately referring to the context or reality of the text phenomena and statements referring to high-level interpretation, transcending the immediate context of the text. A network of interpretative statements may continuously evolve, creating new hypothesis but probably subsequently also demanding new properties and classes.

5.1.3 Humboldt-Universität zu Berlin (HUB)

The third use case stems from the discipline of visual history and the didactics of history. The workshop was held in collaboration with Sabine Moller, from the Department of History of the Humboldt-Universität zu Berlin, who is focusing on Didactics of History. 15 students from the Department of History of the Humboldt-Universität zu Berlin took part in the experiment. All participants had a background in history even though they came from different institutes and participated in different programs (Bachelor and Master). The

[80] In so far, the property should have as its domain and range a class “Topoi”.

experiment was part of the seminar “Fotografie und Geschichte digital” (photography and digital history) which was held over the course of two weeks with four all-day meetings at the end of October and beginning of November 2014. The seminar included presentations by the lecturers on the topic of visual history and historical analysis of historical photographs along with introductions to Pundit and Linked Data.

The particular problem statement of the experiment was two-fold: if Pundit and Linked Data are able to support learning critical analysis of digitised historical photographs and if Linked Data can be used to enable historical critique of visual source documents. In that regard, the seminar was located at the intersection of Didactics of History and Visual History.

During the preparatory phase, we decided to refrain from having the participants create a vocabulary on their own but to prepare a ready-to-use vocabulary for the seminar. The reason was that the translation of the requirements of a historical analysis to a RDF vocabulary demanded more time and effort than would have been feasible for the students in the seminar.

During the workshop the students first learned about how to analyse historical photographs and then were asked to compare what they learned to the prepared vocabulary. The vocabulary was slightly modified and extended during the course of the first workshop day. After a general introduction to Linked Data and exercises with Pundit, the students searched and selected their own photographs from the Web in order to analyse them with Pundit by applying the annotation vocabulary. This phase was attended by the lecturers who answered questions and helped with the functionality of Pundit. Due to technical problems the comparative exploration of the created statements was only possible as a principle demonstration by the lecturers.

The prepared annotation vocabulary constitutes an attempt to translate a methodological approach to the historical and critical analysis of historical images, in this case digitised photographs, to a simple and flat ontology. The method translated was based on several methodological approaches and the expertise of the teacher. In the annotation vocabulary we differentiated the following levels in the analysis of images: (1) the context of provenance including information about the author and the historical context of creation, the shown things in the image (“Bezugsrealität”), the used stylistics in the image (“Bildrealität”), and the historical and personal perception of the image (“Wirkungsrealität”).

During the translation process several conceptual issues arose including the following. The social aspects which need to be considered are potentially unlimited and deciding on relevance on particular aspects in advance is not feasible. The solution was to use generic properties and classes which allow to either create your own textual information (Literals) or by creating your own specific instances, i.e. create your own terminological system. The same is true in the case of existing interpretation offerings and one’s own semantics, where we also resorted to generic annotation entities. For example, we introduced the property “wirkt” (has effect) along with a class for personal and existing interpretative impressions. Here, students were able to create their own instance data with Korbo.

Even though the translation process posed more conceptual obstacles than previous ones, the result nevertheless proved to constitute an applicable and already useful attempt to represent an interpretative approach in the formal and explicit Linked Data structure. The teacher stated also one principal issue with regard to the digital working as a whole that the haptic aspect of the physical image would be lost during and also that spontaneous in interpretative process might not be adequately covered by computers and their formal and explicit working mode.

5.2 Questionnaire

This section summarises the results from the questionnaire for all three experiments. Only where appropriate the discussion will differentiate between the individual use cases. Sections which are specific to the individual use cases have been reported in the previous sections.

The questionnaire was taken on the last day of each workshop and had 31 respondents.

5.2.1 Digital Humanities and Linked Data

The participants were first asked whether they encountered the terms Digital Humanities and Linked (Open) Data before the experiments. 26% (8) of the participants had heard of the term “Digital Humanities” before the experiment while 23 (74%) did not. Only 16% (5) knew of the term “Linked (Open) Data” before the experiment, and 84% (26) did not. Not surprisingly, 48% (15) would not call themselves “Digital Humanist”, and 39% (12) were not sure, while only 13% (4) would say that they are “Digital Humanists”. Accordingly, none of the participants has used a tool for semantic annotations before.

The participants were then asked which advantages they see in Linked (Open) Data tools for their own work. Most participants mentioned the facilitation of information integration, reusability and accessibility of information, i.e. of research data such as the annotations and the research objects and their relation to other objects. Another important advantage seen by the participants as a result of their work with Pundit and Linked Data was the more intensive and different engagement with the research object itself, and that the annotation vocabulary helped to work in a structured and systematic way. The possibility to explore the annotations in Ask was helpful to discover new relations and, at the same time, helped to keep an overview of the annotations created. More generally, some participants pointed out that the terminological system used during the experiment, i.e. the annotation vocabulary and the instance data, creates background knowledge which can be exchanged and reused in working groups.

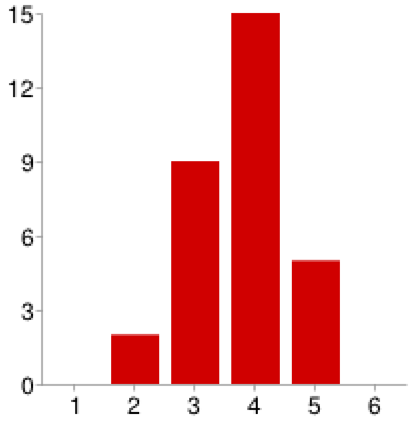
The disadvantages of Linked (Open) Data tools for their own work were mostly related to the specific issues resulting from the current state of Pundit. The issue mentioned most often was the amount of time it takes to create triples, that too many triples result in complexity which is difficult to filter and possible redundancy of statements, and the necessity to communicate with developers to implement new classes and properties. Some participants mentioned the necessity to learn and understand the principles of Linked Data as another potential hindrance.

Participants from the FHP experiment also pointed out that no workflows exist yet, leading again to time consuming work processes, especially that there is no established means and workflows for quality checking before making something publicly available.

Next, the participants were asked in more detail about the functionality provided by Pundit and Linked Data in relation to research in the humanities.

5.2.2 Functionality of Pundit

The participants of the experiments found their experience with Pundit mostly positive, as shown in figure 1. 64% (20) rated their experience as rather positive while 25% (11) rated their experience as rather negative. Considering the current state of Pundit and its components which are not yet optimised for efficient and fluent workflows, this is surprisingly positive.



1	**0**	0%
2	**2**	6%
3	**9**	29%
4	**15**	48%
5	**5**	16%
6	**0**	0%

Figure 10. How would you rate your experience with Pundit? (1 = “very bad”, 6 = “excellent”).

The next question asked for the circumstances under which participants would use Pundit for their own research or work. Mostly, participants demanded general better usability during triple creation, better and intuitive interface in Pundit, and an overall more stable system.[81] These issues related to the current state of Pundit and its components were complemented by requests for additional features such as the option to apply an annotation to several pages at once, for example date or creator to the single pages of a letter, more space to enter free text (Literals), entity extraction, and in particular more and better filter options for annotations. Furthermore, options for access restrictions, for example to Korbo in case of instance data about persons were mentioned as important, and easy export of annotation to other software for further processing. Some participants mentioned support by professionals for creating vocabularies and training with this kind of software and working mode.

The last question in this section inquired about the stages in the research and work progress which participants would like to see supported by Pundit. The participants mostly indicated that one of the main application scenarios relates to researching and collecting facts on one’s own research objects (personal collection), either collaboratively in groups or for individual research. In particular, initial stages of the research process where research objects are pre-analysed and the researcher tries to establish an overview on the corpus, such as formal or outer analysis of sources, appear to be a concrete and immediate application scenarios. The immediate value is seen in structuring and systematising knowledge and to create structured research corpus with connected research objects which also allow to quickly retrieve sources by searching for entities such as persons, places, topics etc. Few participants explicitly mentioned the possibility to use the results from such initial phases for testing a vocabulary for opinion mining applications, or more generally preparing bigger projects and analysis.

The next section focused on various aspects of the digital and non-digital publication behaviour of the participants.

5.2.3 Publication

Roughly half of the respondents indicated to work in rather analogue settings while the rest indicated to work in rather digital settings. Only one respondent said to work only analogue while none said to work only digital. Most respondents are grouped in the middle of the

[81] Servers were down during the HUB experiment for a short time duration, and response times of the annotations servers were lagging occasionally due to the relatively high network traffic caused by the workshops.

scale. These results indicate that genuine digital working settings or contexts are not yet considerably established in the normal working routines of students.

On the other hand, regarding the question whether the respondents would publish their work digitally, 65% (20) answered with yes and 35% (11) answered with no. This indicates that publishing digital is slightly more common than working digitally, i.e. the process of research leading to a publication of research results is less affected by a digital setting.

When asked as to what digital publications are, most respondents provided a broad range of general answers such as any kind of document or information accessible or available online such e-journals, qualification theses via edoc-repositories or websites, Web portals, either as open access or with some access restriction such as pay barrier. Few considered ebooks, music files, or videos files as digital publications as well, while only one named digitised objects available via Europeana or the Deutsche Digitale Bibliothek as digital publications.

The question regarding under which circumstances the participants would publish digitally generated very diverse aspects. All respondents appeared to have considered only traditional text-based publications such as articles here which coincides with the previous responses. Many respondents seem to prefer publication as open access or at least by retaining various rights such as whether one retains the copyright or the right to decide where else the publication will be made available. Few would publish completely freely without any restrictions, while more respondents would allow free access and use limited to academia or particular communities. Other important aspects mentioned were legal consideration regarding copyright and privacy laws in cases of documents related to individuals and the reputation of the publication channel or platform

The annotations created in Pundit were considered as a publication by roughly half of the respondents, 48% (15), while 29% (9) answered no, and 23% (7) were not sure. At the same time, 84% (26) would make their annotations created with Pundit public and available to others. Only 3% (1) would not, and 13% (4) were undecided.

The reasons provided for being willing to make annotations available were manifold. Most reasons were concerned with the potential usefulness for other users. For example, providing annotations on things such as persons or places to others would deliver additional context knowledge on research objects and might help others with search and retrieval. In this context, crowd-sourcing for collecting contextual information on research objects is seen as a kind of fruitful publication. Furthermore, sharing annotations is seen as having the potential to facilitate research and to compare results and to gain feedback in order to improve one’s own annotation data.

However, several concerns and reservations were expressed: Few argued that as long as annotations are not really (re-)useable, for example being citable and referenceable, they cannot be considered to be publications, or that annotations per se are no proper publications and always need support by proper text, that annotations are only supportive to research and more of a collaborative endeavour.

Some would only want to publish “factual” annotations but no “subjective” annotations which are based on interpretative acts or which are personal comments or notes. One respondent made the distinction regarding the content of the annotation: if it is basic, simple information, then it is less important to understand an annotation as a proper publication but if it is more high-level content expressed by the annotation then it is very important. Problematic is also that the reasoning leading to the triple is not obvious and missing which could be problematic. If annotations are considered proper publications then a quality check of annotations would be necessary before publication since the correctness of the annotations is important.

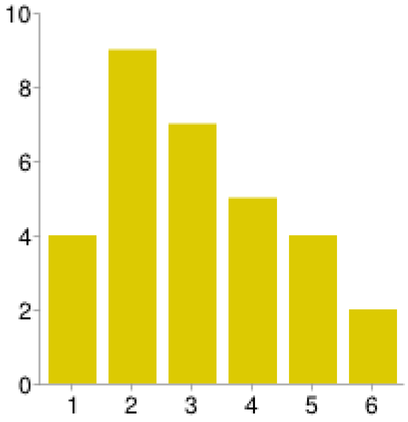
Several respondents raised concerns regarding whether they would retain the rights on their annotations or if potential employer would hold the right on the annotations made as part of a working contract. Uncertainty existed regarding the violation of rights or copyrights of annotated research objects such as digitised photographs or archival material, or personal privacy. Another issue raised was the question what it means to reuse a single entity (resource) from an external knowledge base in a triple: Who would “own” the complete annotation then?

Participants seemed to associate publication of triples as making them accessible in some way but appear not to have pondered about more integrated and contextualised forms of publishing triple data, for example as part of a documented package of statements about a text. This could mean that participants do not consider triples as proper and publishable research. Publication appears to be interpreted as a means for collaboration and supporting each other in research. And so, even though a majority of the respondents would consider annotations as some kind of publication and most respondents would make their annotations available to others, most participants of the experiments were very aware of the potential legal, social, and technical issues surrounding the publication of annotations.

The next section in the questionnaire inquired about annotations.

5.2.4 Annotations

The participants were asked whether they experienced the triple structure of the annotations as restrictive. The responses are spread but the respondent tend to find the triple structure of the annotations as unrestrictive, as shown in figure 11.



1	**4**	13%
2	**9**	29%
3	**7**	23%
4	**5**	16%
5	**4**	13%
6	**2**	6%

Figure 11. Do you find the structure of the annotations (triple) as restrictive? (1 = “I absolutely not agree”, 6 = “I completely agree”.

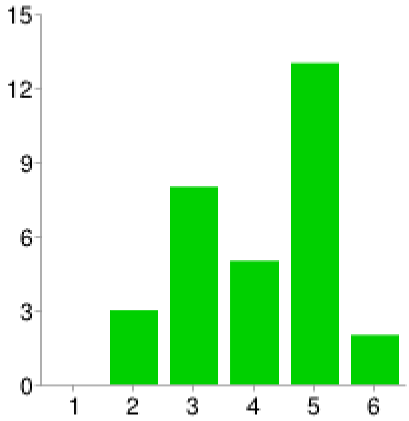
Some of the reasons given for experiencing the triple structure as restrictive include the missing ability for adapting the vocabulary during the course of the annotation work. New statements which are found as necessary during the subsequent work in Pundit cannot be created without extending the existing vocabulary. Pundit, in its current state, does not allow easy extension of the vocabulary because it requires editing and knowledge of configuration files in JSON. In this regard, this is a limitation of the current system. However, in terms of Linked Data principles, extending vocabularies at any time, i.e. not systematically and with rigor and caution, may quickly lead to an influx of statements and schema entities.

Some respondents feared losing information because of missing classes or properties in the vocabulary, or because of the missing possibility to make differentiated statements such as weighted statements. Several respondents criticised that they have to create many triples in order to express slightly more complex information. These participants then also stated that they had to put these background information into free text fields but, at the same time, realised that these free text information are not accessible to further processing.

In general, there appears to be unease and uncertainty towards understanding which annotations are best created as triples and which information is better suited for free text fields (Literals).

However, after having overcome the hurdle of becoming familiar with the principal approach of Linked Data, a majority of the respondents recognised and appreciated the structured approach to annotation as “logical” and systematic. Some advantages stated were that a structured collection of research data evolves and that the annotation vocabulary provides guidance to what should be annotated.

While a slight majority of the respondents tend to feel not restricted by the triple structure of the annotations, a clearer majority considered the provided vocabularies as being adequate for the tasks they had to perform during the experiments (cf. “Use Cases”), as shown in figure 12.



1	**0**	0%
2	**3**	10%
3	**8**	26%
4	**5**	16%
5	**13**	42%
6	**2**	6%

Figure 12. In your opinion, were the provided vocabularies adequate? (1 = “not adequate at all”, 6 = “completely adequate”).

The next question asked how vocabularies should be created and managed: 23% (7) prefer to create and manage the vocabularies on their own, 39% (12) prefer to collaborate with other scientists from the same field, 19% (6) prefer to collaborate with the developers of the tools, and 10% (6) prefer to only re-use existing vocabularies and to leave the creation and management to the developers of the tools.

Various reasons were given in the follow-up question. Most respondents tend to favour a collaborative approach including either several other researchers or additionally developers. On the one hand, the single researchers know their domain and research objects best and therefore know which kinds of statements or extensions they would need. Developers alone would not be able to foresee all relevant classes and properties. On the other hand, however, the danger of losing semantic interoperability of one’s own research data is also seen if the researchers would be allowed to freely manage or successively extend their vocabularies. Therefore, many respondents stressed the importance to collaborate with developers which would help to retain rigor in the vocabulary but also with other scientists in order to avoid too specialised vocabularies or to identify missing entities.

Other responses from FHP also suggested that in larger institutions or working groups, such as archives or divisional departments, selected people could coordinate and manage the vocabulary based on the feedback from researchers in order to retain rigor and also to quickly adapt to new projects. On the other hand, one respondent feared that such collaborative approaches could be too time consuming if larger groups would have to agree on modifications in vocabularies and then depend on implementation by a third party. Being able to add new entities would facilitate the working process.

All in all, the respondents tend to favour collaborative approaches involving researchers and developers to the creation and management of shared vocabularies and stress the importance to be able to specialise their vocabularies in such a context.

The last question inquired for potential reuse scenarios of the personal triples created by the participants. 35% (11) of the respondents could not think of potential reuse scenarios for their triple data. The other respondents mostly indicated three different kinds of principal reuse scenarios: The reuse of the annotations by other researchers working either with the same or similar research objects or on similar research questions. Reuse of previous annotations would be time saving. Some respondents explicitly pointed out that the reuse of annotations, i.e. research data, would be equal to considering previous research. The second principal reuse scenario mentioned by the respondents was using annotations as additional contextual information for search and retrieval allowing, for example, access via person concepts to images. Few respondents pointed visualisation out as a third scenario for further processing the annotations in other programs.

5.2.5 Ontology

In the questionnaire, each use case had a dedicated section with questions tailored towards the particular use case. These questions focused on the application of the specific annotation vocabularies and feedback regarding the principal usefulness of the Linked Data approach for the particular domain covered by the uses case.

HUB

The participants from the Department of History of the Humboldt-Universität zu Berlin mostly stated that they were able to make the most relevant annotations. However, one of the major problems were annotations which would demand to express uncertainty about particular statements, for example, saying that a photograph has been probably taken at a particular time of day. Similarly, expressing assumptions or reflexion, or creating something like a footnote, was not possible but often demanded. Generally, some participants wished for more properties to describe more of the historical background, i.e. context information which is not directly related to the annotated resource itself.

Problematic statements were about the authenticity of photographs, whether a statement about the semantic structure of a photograph relates to the whole image or only a fragment or both.

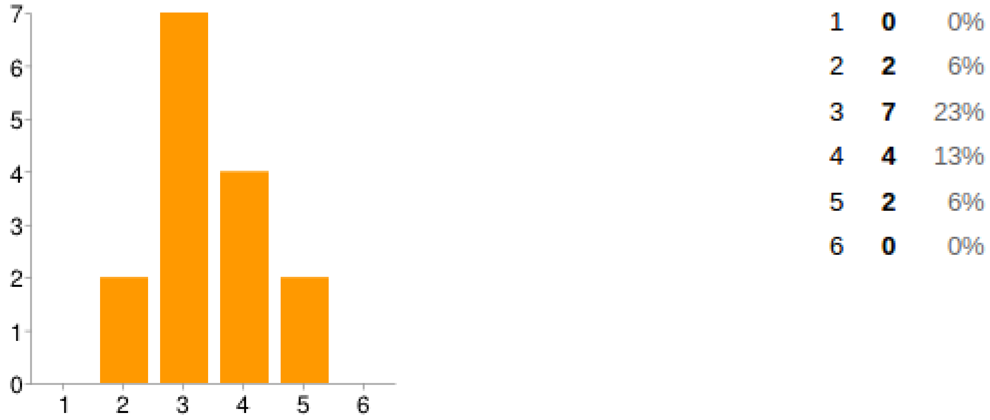
In general, the more interpretation was necessary during the analysis of a photograph (semantic and symbolic structure mostly) the more difficult it became for the participants to reduce these interpretations to factual statements and concepts. Since there was no possibility to express assumptions or uncertainty, many participants chose to not create respective annotations. Furthermore, some participants found the structure of the triple itself as a source for uncertainty because natural language statements lose all grammar.

Some participants stated that the reflection on the process of analysis of a photograph was sharpened by the forced explicitness of the vocabulary and triple structure. In this context, their creativity and inspiration had been aided by the instance data of others available through Korbo but also in Linked Open Data sources.

Apart from comments on the current functionality of Pundit and Ask, which is too time consuming and cumbersome, several remarks were made regarding potential disadvantages for digital critique of images. Some fear to loose information due to the restrictive triple structure. The reason most likely is due to the inability to easily add new properties (and classes) to the vocabularies during the work. Similarly, another mentioned issue is that the interpretative acts are lost in the triple structure, the reason why an annotation has been made, and that, potentially, the creativity of these interpretative acts is lost. Lastly, the annotations do not have any scientific reliability in so far as there are no established measurements for such a purpose.

Some participants had the impression that too much information is being created during the annotation so that the overview quickly diminished. This impression is certainly due to the limited facilities of Pundit and Ask to easily filter annotations. Being able to freely create triples might also entail the danger to lose focus on the actual objective of the current working task because you can go on with triples “in any direction”. Other feared that too many allowed and publicly available perspectives and opinions - expressed through annotations and instances created by users - could lead to a lot of wrong or bad information.

The final question in this section inquired whether the Department of History of Humboldt-Universität zu Berlin participants think they overall successfully worked with Pundit and the annotation vocabulary on the research questions and tasks. As figure 13 shows, respondents tended to judge the success more sceptical than the participants in the FHP experiments which will be discussed next. However, considering the complex and difficult topic operationalised for this use case, the feedback can be considered as encouraging.



1	**0**	0%
2	**2**	6%
3	**7**	23%
4	**4**	13%
5	**2**	6%
6	**0**	0%

Figure 13. Would you say that you successfully worked on the research questions and work tasks with Pundit and the provided ontology (1 = “not successful at all”, 6 = “completely successful”)?

FHP

In the case of the FHP experiment, most participants were able to create all relevant statements. Examples of missing statements include provenance relations, i.e. relation between document and holding or collection, and the relation to the archive, structural relations such as next page, and various details such as nicknames of persons, gender, additional information about places, or a dedicated property for describing the content of a text. Nearly all of these shortcomings were solved by the participants by using the free comment property. All specified missing statements would be easy to provide from existing ontologies. In the context of the experiment, time constraints prohibited adding appropriate entities to the vocabulary.

Most participants indicated that they were able to make all statements they deemed necessary during the workshops. One participant specifically asked for the possibility to create class hierarchies in order to allow clearer ontological differentiation between class and instance. Few participants felt that the German labels for the properties in particular were not always fitting in terms of the articles. Some asked for better class and property descriptions.

Some participants missed an option to create complex statements, for example by combining several triples. The facility of Pundit to create templates for RDF statements was not used in the experiments but might have been a possibility to address this issue. Few specific, conceptual issues were raised such as how to deal with the changing borders and the location of historical places. The target of annotations in the context of specific statements was not always clear to the participants. For example, should place names or person names be annotated by using the complete document or the particular text-fragment as the subject of the triple. Another general question was where the line should be drawn between the information the archivist and the information the researcher should annotate.

As advantages the often mentioned aspects were stated such as collaborative work, additional context for documents, the re-usability of annotation data, and connecting and integrating information.

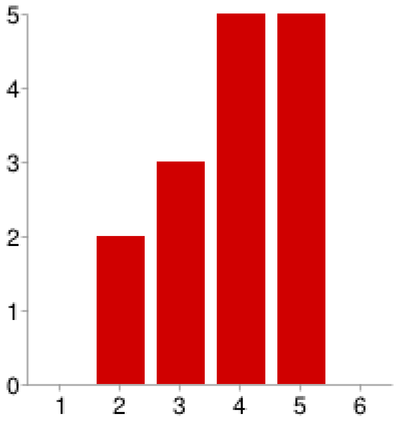
Potential disadvantages were that, at least in the context of Pundit, every annotation can be edited or deleted leading to potentially unstable and unreliable research data, i.e. the annotations. Furthermore, there is no reliable system for checking and sustaining the integrity and authenticity of statements. A related question is how trust in terms of the scientific reliability of the Linked Data sources could be established.

Even though collaboration has often been mentioned as positive, i.e. in terms of sharing the workload or re-using information, some participants also pointed out that collaboration may also lead to confusion and potentially dissent during work. Also, if collaboration is understood as involving people from outside a group or organisation, this could mean losing control over the documentation process if anyone would be allowed to annotate.

Some respondents stressed the importance of having stable and accepted vocabularies in order to have relevant annotations and avoiding too much information. Achieving this aim, however, demands a lot of work and coordination before productive annotation could be done.

Also, the necessary training regarding Linked Data principles and tools and the possible additional work in terms of coordination and quality control have been mentioned as potential obstacles.

FHP participants were also asked how they would judge the success they had in translating and applying the editorial guideline to Linked Data and Pundit. The responses are fairly positive, as shown in figure 14, which confirms the general tendency of the responses provided in the other questions.



1	**0**	0%
2	**2**	6%
3	**3**	10%
4	**5**	16%
5	**5**	16%
6	**0**	0%

Figure 14. Would you say that the translation and application of the editorial guideline into the Linked Data format was successful (1 = “not successful at all”, 2 = “completely successful”)?

FHP respondents were further asked about potential usage scenarios for Linked Data in the context of digital editions and archives. Most respondents said that Linked Data could be used, in principle, for guidelines for digital editions but also stressed that proper vocabularies, documentation, and workflows are necessary, and that the actual tools need to be easier to use.

In the case of archives, most respondents were sceptical regarding the application of Linked Data and annotations with tools such as Pundit in archival contexts and especially the daily work of archives. The Linked Data approach in combination with annotations may have potential for (explorative) search functionality in archives allowing to discover sources connected to particular entities, or, more general, to provide additional contextual information on particular documents or holdings.

The biggest obstacle, however, is the lack of resources in archives to learn and employ such techniques and tools, the mass of documents, and the issue of access restricted holdings. Regarding the latter, participants called for proper technical and administrative policies which would ensure access restriction to triples describing such holdings. Therefore, in the short-term, the Linked Data approach may be mostly useful in the context of small and closed projects, for example for presenting and publishing collections of documents about a topic, but not for daily archival work. In the future, the Linked Data approach could potentially be used for the description stage in the archival workflow where information such as call numbers or content description are added to archival material.

Lastly, in the case of editing source documents, most respondents were also relatively sceptical. A principal advantage mentioned is the flexible terminological system consisting of instances created within a structured and stable framework, the vocabulary. The workshop showed that the annotation already worked well for small comments and editing small remarks, but is less useful for longer transcriptions. In general, some respondents felt that the annotations are too detached from the edited document.

GEI

The GEI participant first provided feedback on the annotation vocabulary and the annotation process.

During the work process, additional properties seemed to be necessary. During the interpretative work with the school books several layers of the interpretation became apparent. Statements differ in their level of specificity and how directly they relate to phenomena which are immanent to the text. These different layers should be incorporated into the annotation vocabulary in order to differentiate them. Most problematic were statements where the participant was uncertain about the statement: Differentiating how sure he was about a statement was not possible but deemed very important.

However, the participant was not sure whether new properties should be introduced due to potential conceptual and technical issues such as too many properties which could diminish re-usability. The respondent stressed that sufficient time for preparation is very important to prepare a stable and conclusive vocabulary. However, all in all, the vocabulary proved to be useful and sufficient for most relevant statements.

The participant would refrain from creating annotation vocabularies alone but would prefer to do so in collaboration with developers. The Linked Data and Semantic Web principles, for example, is not easy to understand and the participant would prefer to have explanation for important terms provided by experts.

Furthermore, the possibility to create hierarchies of classes and properties is important in order to create a proper terminological system for annotation. Here, the current functionality of Pundit prohibited to create class taxonomies for the annotation vocabulary, however, RDFS does provide the necessary means for building such taxonomies.

Lastly, the participant would have preferred to have more options for a more flexible visualisation of the created statements. The respondent preferred a chronological visualisation of the annotations. Even though the faceted browser allowed to explore the statements created, a possibility to have a more intuitive chronological representation of the statements would have been preferred. For example, showing all triples including a connotation chronological and in the context of the schoolbook the annotation was made in.

Regarding the potential usefulness of Pundit and Linked Data approaches, the participant identified as potential useful application scenarios of Linked Data and Pundit collaboration within research groups and projects where work task could be easily assigned to different people and the results then merged afterwards. For example, assistants could be given the annotation vocabulary as a guideline and catalogue of criteria by which they would search and describe documents in an archive such as historical journals or newspapers. The annotations created by using the vocabulary would allow assessing, comparing, and merging the results from the different assistants more easily.

In this context, another advantage was seen in the fact that the annotations vocabulary and the created instance data, the annotations, are independent from the application. However, Pundit is missing an easy option to export this data in order to reuse it in different contexts. For example, other applications could reuse the created statements and annotation vocabulary as a basis for automated analysis of a large corpus of schoolbooks.

Major problems and disadvantages of Pundit and Linked Data were, according to the participant, the relative high effort necessary to create annotations. This process is too time consuming at the moment in terms of the usability of the annotation tool Pundit, where it takes too many clicks to create an annotation.

There was unease with the technical infrastructure on which the participant felt dependent in so far as he would not be able to work anymore if the server or the Internet connection broke down. The participant would prefer some kind of backup or intermediate or local temporary storage which would allow to work at any time and independently from a working internet connection or server. Changes would be synced back to the server when being online again.

Another potential problem pointed out are legal issues. The data used for this use case is under control of the GEI. However, in general, the participant stated that it is unclear to him how exactly copyright and reuse of digital sources, digitised documents or Linked Data instances, are handled and what that means for one's own research work.

The GEI participant judged the general success of the work conducted in the context of Pundit and Linked Data as more positive (scale of 4).

5.3 Conclusion

The experiments were conducted within the limitations imposed by the current state of Pundit which primarily include the inability to create class taxonomies in annotation vocabularies and the limitation to faceted browsing in tabular formats based on the entities incorporated in the vocabulary and instance data (schema and instance elements). Despite these limitations, the results of the experiments underline the usefulness of the principal approach of Pundit and of Linked Data for genuine research questions and interests of humanists by providing evidence for their applicability in the context of interpretative approaches in the humanities.

In each use case, we were able to create useable drafts of annotation vocabularies in RDF(S) for the respective research interests in relatively short time: for the analysis of connotations in historical schoolbooks, the historical critique and analysis of photographs, and for editorial work on archival documents. Each annotation vocabulary remained simple and without complex ontological constructs. Even though several interpretative processes were difficult to explicate and even more difficult to formalise, the iterative translation process is already a genuine part of the interpretative research process. This iterative process of translating methodological approaches and interpretative statements, not only during preparatory stages, but continuously during the actual annotation phase, needs to be considered much more intensively and systematically than was possible in the context of these experiments. New and necessary statements will appear only during the application of the annotation vocabulary and need to be fed back into the vocabulary. For constructs which were not representable in the annotation vocabularies, informal conventions have been introduced, for example, by prescribing particular textual values for objects in triples in cases where information was either unknown or uncertain.

However, all annotation vocabularies proved to be reasonably productive for their respective purposes. In this regard, simple annotation vocabulary in RDF(S) appear to be able to support very different kinds of research interests, in the context of the experiment, that is the archival-historical domain.

In particular, simple Linked Data annotation vocabularies proved to be relatively accessible to humanists. Even though all participants had no prior working knowledge of Linked Data or Pundit, they were mostly able to grasp the concept of triple annotations within a couple of hours and thereafter apply the annotations vocabularies. In this regard, results appear to be obtainable for students with relatively low prerequisites which is an important aspect for lowering the access barrier to Linked Data annotation tools.

Furthermore, the formal and explicit approach of Linked Data appears to have initiated reflection on participants" own working practices. Student were forced to reflect on their own work processes because of the explicitness of the vocabulary. The annotation vocabularies and the created instance data provide a common basis for discussion on the method, what should be said, and interpretation, what has been said, of the research objects in particular contexts since Korbo and faceted browsers allow to explore any triple data relatively easily. On the other hand, teachers have a pedagogical tool for communicating theoretical and practical representations of methods. In so far, the annotation vocabularies and Pundit constitute a potential epistemological tool for educational contexts.

The reuse of the created annotation vocabularies in other similar use cases remains open since much work would have to be invested into their further specification in order to be useable by other researchers. However, the experiments could also be seen as a first round of evaluation and testing of such vocabularies providing the basis for future refinements. In order to be able to develop "real" application scenarios a translation between the requirements of the researchers and their respective research process, on the one hand, and the functionalities of the virtual research environment developed for them, on the other hand, must happen in an iterative translation process.

Future experiments with Pundit or similar tools would need to prepare more specific adaptations to the single use cases, especially in terms of flexible visualisation and filter options which proved to be the most pivotal incentive and means for establishing a sense of usefulness for proper and sustained research. Chronological visualisations and immediate contexts are important and comparison of entities.

The experiments resulted in several recommendations regarding the general research question of Task 3.4 "How can Linked Data based digital tools and data support, facilitate, or enhance humanist work practices?" The following list contains the most relevant recommendations:

1. Providing clear mission statements towards the purpose of a digital tool and vocabulary in relation to the overall research process and particular stages is essential. For example, which stage or segment of the research process does an annotation vocabulary address? The general point has been made in the context of the SDM already: One of the most important aspects is communication between the humanist and the developers.

2. In this regard, the aspect of good usability of tools and interface needs to be stressed again. Without ease of use and efficiency, tools such as Pundit will not be used voluntarily in any serious or productive scenario. Tools need to provide feedback at any point why a particular functionality does not work.

3. Tools and workflows need to be established, in order to implement and maintain a collaborative and iterative development and application of the annotation vocabularies for interpretative approaches. This includes a vocabulary browser and editor which allows to edit and extend vocabularies by classes, properties and scope notes in a controlled and flexible manner. How appropriate policies and workflows should look like remains subject to future research.

4. For interpretative applications of annotations vocabularies it is important to investigate how to qualify statements, for example, allowing to provide reasons or to express uncertainty, and to ensure proper provenance information. Interpretative statements are the result of proper research work and need to be attributable and citable. Furthermore, access rights management for statements is important where access to statements can be regulated granularly. These are difficult topics which need to be addressed, however, in order to start establishing scientific reliability and trust.

5. The stability of the annotated research objects and instance data is another crucial aspect. If the annotated objects or instance data used in triples from Linked Open Data knowledge bases disappear, the result of the interpretative research work is rendered invalid. This issue has been raised by participants. Furthermore, the teachers pointed out that they have to grade the students also based on the triple statements they created. Here, sustainability of these triples is crucial and is a clear provenance of statements.

6. In this context, the option to export the research data, i.e. the created annotations and instance data from the notebooks to locally stored files is important. Even though not all participants raised the issue where annotations are currently stored, or were even conscious of the problem, when directly asked about the issue, the need to export data was stressed. Motivations were to reuse the data in other applications such as for more sophisticated visualisations, to have full control of the data, and to have a backup of the data available. This will contribute to more trust in digital tools regarding the safety of the personal research data.

7. Lastly, legal issues of using and reusing Web resources of any kind including Linked Data resources, need to be discussed, explained and communicated. Even though users of the Web are aware that the same or similar legal constraints and formal and informal obligations pertain to the use of digital resources as in the analogue world, uncertainty about the exact ramifications, rules, and regulations prevail. Several examples have been given in the previous discussion pertaining to the use of images but also to Linked Data resources such as the legal status and intellectual rights towards annotations and even single resources, whether thumbnails are already duplications of the original picture, or the “Schöpfungshöhe” of annotations. The legal status of these matter to humanists.

8. Approach the students and young researchers outside the “Digital Humanities”!

For future research and further development of workflows for interpretative approaches implemented with semantic annotations tools such as Pundit, we suggest the following 3-tiered process. Even though partially predetermined by Pundit, the three steps proved to be useful and an appropriate approach to engage humanists with Linked Data in a way which, provided appropriate implementation of the necessary tools, will gradually allow to improve the usefulness for humanist research. The same basic iterative 3-tiered process appeared to be valid in the context of the reasoning experiment (cf. “Reasoning”) where interpretative modelling has been investigated in the context of Pundit and Linked Data.

We therefore propose an iterative 3-tiered process to be implemented and offered to humanists in order to enable them to begin with meaningful and useful interpretative work with Linked Data enable semantic annotation environments:

1. *Conceptualising*: Selecting, modifying or creating a vocabulary (referential structure) for annotation is already a genuine part of the research process. This process needs to involve both sides, the humanists and developers. A method and policy to revisit and modify the initial vocabulary needs to be implemented since necessary statements develop during research work, the annotation work.

2. *Annotating*: Applying the annotation vocabulary to a research object is the second step. During this phase, humanists appear to prefer to create their own referential data, either because necessary instance data does not exist or is not being trusted (the Linked Data knowledge bases appear alien to the participants). If they import instance data they prefer to have a filter on import in order to have control over what is being imported into their notebook. Important is also to allow statements regarding the rationale and weighting of a statement which was one of the major demanded features.

3. *Visualising*: The process of exploring what has been created in terms of the annotation vocabulary, statements and instance data. Here, humanists want to apply their own “reasoning” by filtering and adopting the visualisation context to their needs. Relevant generic visualisation were the simple comparison of two or more entities of the same type and their immediate attributes, and chronological ordering and displaying of statements.

Acknowledgments

The following members and associates of the Digital Humanities Advisory Board (DHAB) contributed to the development of the Scholarly Domain Model and the writing of the paper: Markus Schnöpf (BBAW/FHP), Andreas Weiß (GEI), Sabine Moller (HUB).

6 Report on Reasoning

The use and application of digital research environments is of growing importance in the humanities. Within the discipline that has emerged out of the joining of the two fields of humanities and computing, the Digital Humanities (cf. Svensson 2009), there is an ever growing number of projects embracing Semantic Web technologies and Linked Data especially. As with all Digital Humanities endeavours, the question arises as to what extent the technologies developed in the context of computer science translate to the actual requirements of scholars in individual humanities disciplines (cf. McCarty 2005: 141). In 2009, Zöllner-Weber discussed the specific topics of logic reasoning and ontologies for the humanities. In her study regarding the use of inference tools in the domain of literature studies, she came to the conclusion that there are limitations to the application of such tools for humanities scholars. Not only does the use of these tools often require an in-depth understanding of mathematical logics, but the traditional scholarly activities in the humanities often involve "vague, ambiguous, or even contradictory" (Zöllner-Weber 2009: 10) information. In this sense, McCarty argued in 2005 that the benefits of computer science, which "focuses on combinatorics, syntax, and algorithms" and whose "guiding question is "*what can be automated?*" fail to "address the humanities intellectually." (McCarty 2005: 141) This leads us to the question dealt with in this paper: What kinds of "reasoning" can humanists in fact apply with benefit to digital data, in particular, to Linked (Open) Data?

When we talk about "reasoning" in the context of the Semantic Web and the Digital Humanities we have to consider two principal senses of the term: the algorithmic use of the term as machine-supported inference of new knowledge, i.e. the creation of new relations in the graph, from a given knowledge base and the use of the term as the way humans in general, and humanists[82] in particular, apply their styles of reasoning to the data and which inferences they draw (cf. Blanke et. al. 2013).

The first sense appears to be the most prominent interpretation and topic in Semantic Web digital humanities research[83] and often seems to obscure the second one. Semantic Web reasoning understood as large-scale machine-based inference, however, is not always accessible, feasible or even appropriate for applications and research questions within the digital humanities. All too often, the algorithmic potential of computers blocks the view of the seemingly simple but functionally useful Semantic Web tools available to the scholars already at present (cf. McCarty 2005).

The focus of research needs to be more inclusive with regard to the second sense and to examine if and how Semantic Web tools can support practices of reasoning and thinking about research topics typical for the humanities. This could necessarily build the basis for the application of reasoners down the road, but should also, and in the case of this deliverable primarily, serve to elucidate how scholars can work with such existing tools in the short term. According to McCarty (2005), "in the world of computational things we tend to value intricate, complex, algorithmically sophisticated tools, and so to undervalue what we actually have (...), [the] crude but functional" applications which allow us to explore new potentials (McCarty 2005: 112-113). These new potentials are arguably found in the data (cf. Oldman et al. n.d., or Gradmann 2013b) but also in the application of reasoning and problem-solving abilities of the human mind. In this sense, Deegan/Sutherland (2009) said the following in the preface to "Transferred Illusions. Digital Technology and the Forms of Print":

[82] We use this term as a translation of the German word for "Geisteswissenschaftler" and not in the political sense.
[83] The topics dealt with in the International Summer School's yearly edition of "Reasoning Web" (z.B.2005-2012) are a testament to the prevalence of this notion of reasoning within the digital humanities.

"As books do, computers measure, store and represent information; and as with books, it is still authors and readers who process this information as knowledge, making sense or nonsense of what the tool communicates. So far, computers cannot replicate common sense or creative thinking. The intelligence involved in processing information into knowledge is only a human skill."

"Human reasoning" can therefore be seen as an alternative to computer reasoning and a prerequisite for what McCarty calls "human computing": "Human computing is computing that only proceeds with the essential and continuous engagement of human cognitive processes." (McCarty 2005: 147)

In this deliverable we, from a humanist point of view, will look at the application of human reasoning assisted by relatively simple digital tools, in particular tools to collaboratively and intellectually create and query Linked Data. In this context, we will specifically focus on the tools for the collaborative semantic annotation of digital resources that have been developed by Net7.[84] These include Pundit[85] (Grassi et al. 2013) and its family of applications: Korbo[86], Ask[87] and its built-in faceted browser for querying the semantic annotations made with Pundit. We will use these tools, in particular faceted browsers, as the basis for experiments being conducted with humanities scholars at two DM2E partners: the Wittgenstein Archives at the University of Bergen (WAB), [88] and the Georg-Eckert-Institute for International Textbook Research (GEI).[89] These experiments should shed light on how these tools may support, facilitate, or even hinder humanist reasoning in a digital research environment based on Linked Data. Furthermore, the term "interactive reasoning" may characterise the practices that arise at the intersection between humanist reasoning and the Semantic Web reasoning by stressing the active involvement of the researcher during reasoning processes, i.e. how humanist researchers use Linked Data, or any data in a digital setting for that matter, to come to conclusions and find meaning with regards to their research questions. In our specific case, we will be focusing on a particular example of such interactive reasoning, namely faceted browsing. The aim of these experiments is not to achieve a systematic overview of all types of humanist reasoning that can be associated with Linked Data tools, but to investigate the way in which particular researchers may use their own styles of reasoning, typical of the humanities, to engage with Linked Data utilising simple exploration tools such as faceted browsers. We thereby strive to contribute a different perspective on the Semantic Web reasoning discourse.

First, we will introduce the context of reasoning within the Semantic Web domain. Then, we will discuss the term "humanist reasoning" using the work of Holyoak and Morrison (2012), McCarty (2006) and Hacking (1985) as a basis. Afterwards, we will explore humanist reasoning use cases with DM2E partners at WAB, and GEI who are working with the DM2E tools Pundit and Ask. Finally, based on the observations that arise from these use cases, we will discuss potential Semantic Web reasoning applications in the first, computer-aided sense.

[84] Net7 (http://www.netseven.it/en/) is the leader of WP3 in the DM2E project (http://dm2e.eu/).
[85] https://thepund.it/
[86] http://www.korbo.org/
[87] http://ask.thepund.it/
[88] http://wab.uib.no/, with an open access edition of a part of Wittenstein's Nachlass hosted at http://wittgensteinsource.org/ and host of the Open Access Wittgenstein datasets http://wittgensteinsource.org/ (primary sources) and http://www.wittgensteinrepository.org/ (secondary sources).
[89] http://www.gei.de/en/home.html

6.1 Reasoning

In this section, we will first discuss what the term “reasoning” traditionally means in the context of the Semantic Web and why its implementation in the humanities is difficult. We will then discuss what “reasoning” for scholarship on the basis of Linked Data in the humanities entails and explain our concept of “interactive reasoning” as a practical and complementary alternative to the concept of automated “reasoning” in the Semantic Web.

6.1.1 Semantic Web “Reasoning” [90]

In the Semantic Web, the term “reasoning” generally describes the ability for machines, so-called “reasoners”, to automatically draw inferences from certain types of prepared data using formal logic and Description Logic; work in this area is related to the field of knowledge engineering (cf. Ludwig 2010). For this purpose, data is formalised in a triple structure based on the RDF data model. The semantics of that data is described by classes and properties which are formalised in ontologies. One simple way to describe the purpose of reasoning is that it is for “discovering new relationships”[91] in the existing data, this will be of importance when discussing how humans can interact with Linked Data in later sections of this deliverable. Here, we will discuss in more detail the basic elements required for Semantic Web Reasoning, which also ultimately play a role in understanding the “interactive reasoning” with Linked Data proposed here. These elements include RDF triples, RDFS, OWL, SPARQL, vocabularies and ontologies.

Linked Data Concepts

Semantic annotations according to the Linked Data paradigm at the most basic level consist of RDF-triples, which are simple statements about (Web) resources using an abstract syntax that is human and machine readable. This simple structure is analogous to basic sentence formation in natural language (especially English) and consists of a subject, a predicate and an object, where, according to the World Wide Web Consortium’s (W3C’s)[92] RDF-Primer, “[t]he subject and the object represent the two resources being related and the predicate represents the nature of their relationship”.[93] Web resources as well as the relations connecting them are named with a Uniform Resource Identifier (URI), commonly in the form of an HTTP Uniform Resource Locator (URL) so that they can be unambiguously identified, easily found on the Web, and reused by other scholars.[94]

The following example shows the elements needed to create an RDF triple stating that a certain “resource on the Web” (subject), here from a text published in Wittgenstein’s Nachlass on Wittgenstein Source, “discusses” (predicate/relation) the philosopher “Plato” (object). Plato is uniquely identified in this example using the URL from the corresponding DBPedia.org page.[95] The relation is uniquely identified with a persistent locator that has been catalogued and registered with purl.org.[96]

[90] Readers familiar with Semantic Web “Reasoning” may skip to the next heading. An introduction has been included here, since this chapter of the Deliverable will be published as an article.
[91] http://www.w3.org/standards/semanticweb/inference
[92] http://www.w3.org/
[93] http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/
[94] http://www.w3.org/DesignIssues/LinkedData.html
[95] http://www.dbpedia.org/
[96] https://purl.org/docs/index.html

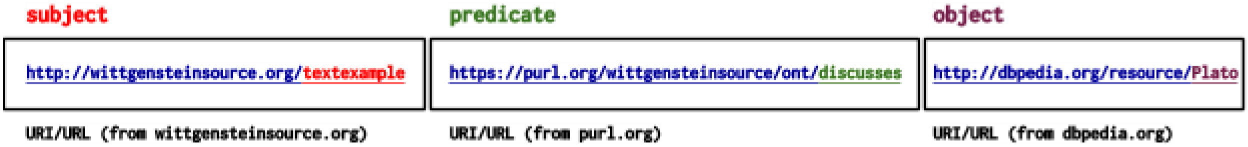

subject
predicate
object
http://wittgensteinsource.org/textexample
https://purl.org/wittgensteinsource/ont/discusses
http://dbpedia.org/resource/Plato
URI/URL (from wittgensteinsource.org)
URI/URL (from purl.org)
URI/URL (from dbpedia.org)

Figure 13. Basic triple structure: Subject - Predicate – Object.

Triples create a graph structure that can be infinitely extended by connecting nodes using new relations, making, for example, an object of one triple the subject of a new triple. The graph is often visualised as follows:

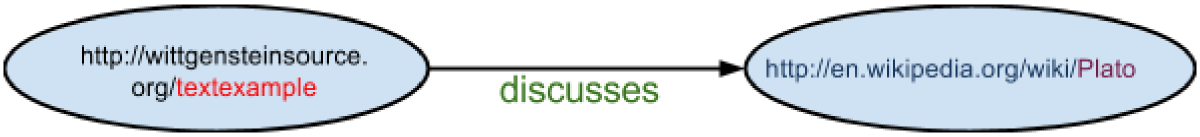

http://wittgensteinsource.org/textexample
discusses
http://en.wikipedia.org/wiki/Plato

Figure 14. RDF triple as a graph.

The power of the graph can perhaps best be demonstrated by the ubiquitous Linking Open Data cloud diagram, which "shows datasets that have been published in Linked Data format,[97] by contributors to the Linking Open Data community project[98] and other individuals and organisations."[99]

[97] http://linkeddata.org/

[98] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[99] Note that the DM2E data is also part of the LOD cloud (http://lod-cloud.net/).

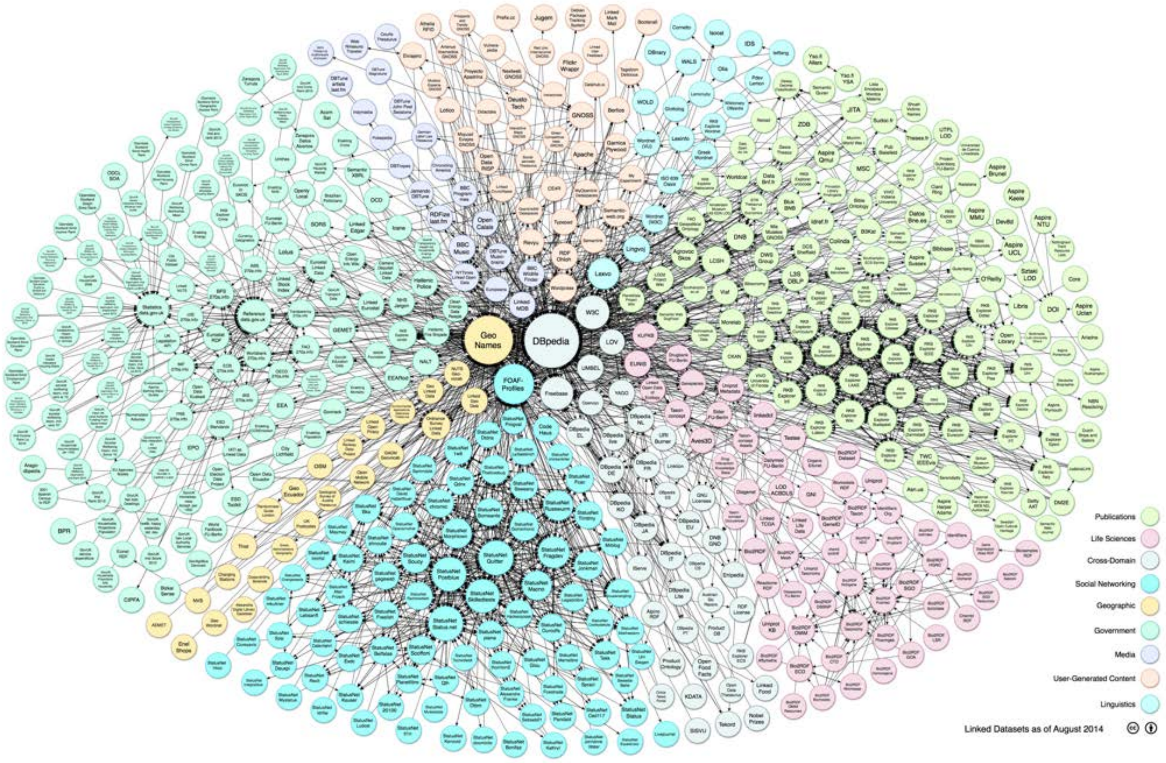


Figure 15. The LOD diagram from 2014.

The RDF Data Model is an abstract syntax[100]. In order for it to be useful for machines and humans in the modelling of data, it not only needs to be formalised as a concrete syntax (the exact rules for writing and storing the triples), but there also needs to be some consensus about the meaning of the predicates used and how they represent the relationship between the subjects and objects. For the former, RDF vocabulary and RDF syntax languages such as Turtle[101] or RDF/XML[102] are used; they will not be considered in detail in this deliverable. For the latter, schemas, vocabularies and "ontologies" have been established.

The Resource Description Framework Schema (RDFS) extends RDF by providing "mechanisms for describing groups of related resources and the relationships between these resources."[103] More specifically, RDFS provides a vocabulary for defining classes and properties, and to create subclass and sub-property taxonomies. Furthermore, the domain and range of properties can be specified. These constructs have simple predefined semantics which already allow simple kinds of reasoning such as transitive reasoning along subclass relations. RDFS therefore provides a basic "grammar" (Gradmann 2013) for the semantic modelling of data, it cannot, however, cover more complex modelling needs. For this

[100] http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/
[101] http://www.w3.org/TR/2011/WD-turtle-20110809/
[102] http://www.w3.org/TR/REC-rdf-syntax/
[103] http://www.w3.org/TR/rdf-schema/

purpose, there is the Web Ontology Language (OWL),[104] which is now in its second version as OWL2.[105] For the purposes of this deliverable, we will use the term OWL to refer to both editions.

OWL and its sublanguages (OWL Lite, OWL DL and OWL Full) facilitate the development of ontologies "by providing additional vocabulary along with a formal semantics" based on Description Logic (DL) for this purpose.[106] Ontologies created on the basis of OWL not only help to structure the knowledge in a certain domain, they also allow for the inclusion of more logical constructs which can then be "understood" and processed by machines. They therefore serve as the basis for machines to complete reasoning tasks. Gruber (1993) describes the term ontology as "an explicit specification of a conceptualization." In other words, creating an ontology is a way to represent and contextualise a certain section of reality (cf. Gradmann 2013: 222).[107] One could also say that the ontology helps to create a knowledge base that "store[s] knowledge about the real world" in a certain domain (cf. Ludwig 2010).

One common way to query the data stored as triples is by using the W3C recommended querying language SPARQL,[108] an acronym of "SPARQL Protocol And RDF Query Language", in connection with a SPARQL endpoint interface. Accessing the data in this manner demands not only a knowledge of the language itself, but also previous knowledge of the types of entities and relations in the triple store as well, which can be quite a barrier for those not familiar with the dataset when trying to query the data.

Reasoning and Inference using Machines

One major area of interest in the Semantic Web domain is being able to process the formalised "representation[s] of terms and their interrelationships"[109] expressed using ontologies by applications so that logical inferences about the data can be made on a large scale. This is, as mentioned above, generally considered "reasoning" in the Semantic Web sense and is geared towards automatically finding implicit knowledge in the data. Reasoners for automatically finding such logical connections inherent in the data created have to be tailored to the particular ontologies and needs of the specific research question and domain (cf. Gardiner 2006). In their "Comparisons of Reasoners for large Ontologies", Dentler et al. (2011) provide a solid definition of the term "Reasoner": "A reasoner is a program that infers logical consequences from a set of explicitly asserted facts or axioms and typically provides automated support for reasoning tasks such as classification, debugging and querying."

As mentioned above, a reasoner uses logic based in mathematical theory to infer new information automatically from the existing triples in the graph. The W3C provides a simple example of how this type of inference works; note its similarity to the philosophical syllogism:

"The data set to be considered may include the relationship (Flipper isA Dolphin). An ontology may declare that "every Dolphin is also a Mammal". That means that a Semantic Web program understanding the notion of "X is also Y" can add the statement (Flipper isA Mammal) to the set of relationships, although that was not part of the original data. One can also say that the new relationship was "discovered". Another example is to express that fact that "if two persons have the same name, home page, and email address, then they

[104] http://www.w3.org/TR/owl-ref/
[105] http://www.w3.org/TR/owl2-overview/
[106] http://www.w3.org/TR/owl-features/
[107] For example http://www.w3.org/TR/owl2-primer/#Class_Disjointness
[108] http://www.w3.org/TR/rdf-sparql-query/
[109] http://www.w3.org/TR/owl-features/

are identical". In this case, the "identity" of two resources can be discovered via inferencing."[110]

This type of reasoning relies primarily on the consistency of the outward form of the statements in the dataset, which not only have to be apophantic, but also have to accurately represent the objective reality of a certain domain on some level to provide the conditions for relevant conclusions (cf. Zoglauer 2008). The original statements also have to have been at some time provided by humans based on their prior knowledge and reasoning about the domain. A Semantic Web reasoner cannot, for example, interpret whether the underlying information (presuppositions) that has been represented in the rule set or vocabulary is actually factual, valuable, sensible, objective, unbiased, relevant or even useful. Only a human has the ability to acquire knowledge and determine meaning. Therefore, a dataset could state, for example, that "Flipper isA Dolphin" and the ontology that "Every Dolphin is also a Bird". The resulting information "Flipper isA Bird" deduced from the initial premises would be logically coherent (cf. Zoglauer 2008: 9), but for the human observer of course mere nonsense. Humanists, however, are not necessarily interested in the form, but in the meaning of statements (cf. Oldman et al. n.d.) about objects in their domains.

This type of Semantic Web reasoning can be used, according to the W3C,[111] for "improving the quality of data integration on the Web", or may help the researcher to "automatically analyse the content of the data" or to "discover possible inconsistencies in the (integrated) data". In certain circumstances it may even be used for "discovering new relationships". This process, however, is contingent on facts being explicitly stated by the scholars and is therefore limited to their ability and willingness to do this, the socio-historical context in which they do this, and the quality of the information contained in the modelled data. Such uses for Semantic Web reasoning are therefore, in the short term, not necessarily as promising for humanist research using Linked Data, not in the least because this type of reasoning is limited in scope.

Difficulties of Semantic Web Reasoning in the Humanities

There are several reasons why this type of machine-supported inference has limited use and relevance from the perspective of a humanities scholar. We will mention three obvious ones here.

First, understanding and utilising this type of machine-aided reasoning requires at the very least a basic knowledge of concepts and skills that are uncommon in most humanities domains. They include, but are not limited to computer programming and querying languages, Linked Data concepts such as the ones previously introduced, database management, ontology creation and knowledge representation, the use and implementation of inference machines, and formal logic in its mathematical expressions such as Description Logic. Even power users including the Digital Humanist might have a steep learning curve for some of these skills. Persuading a humanist to take the time to learn how to apply Semantic Web reasoning requires at the same time a clear understanding of the benefits that will be reaped for her domain and specific research interest.

Second, the objects of study, the types of research questions, and the methods found in the humanities are not always compatible with the Semantic Web reasoning paradigm. As mentioned above, mathematical logic is mainly concerned with the form of statements while, in contrast, humanists have complex and often contradictory research objects (cf. Oldmann et al. n.d.) and are interested in layers of meaning. For example, a historian has little use for creating labour intensive knowledge representations that allow a computer to

[110] http://www.w3.org/standards/semanticweb/inference
[111] http://www.w3.org/standards/semanticweb/inference

“infer” that Flipper – a fictional character from a TV-Show – is a mammal. Unambiguously defining an object is rather the concern of the applications of natural sciences, where such inference machines have been successfully implemented. The historian is however perhaps more concerned with what this fictional character might represent to the audience of one or more time periods. This requires extensive knowledge about several domains such as the culture of the society in general and television in particular, the history and culture of the reception of the show, the particular language of the imagery used in the series and its relationship to other shows. The information (datasets and ontologies) required for an algorithm to automatically come to the historian’s conclusion would be difficult to create and implement. A conclusion the historian might draw such as “Flipper isA(n) aquatic Lassie” is metaphorical, highly subjective, and neither true nor false, making it not a necessarily good candidate for the premise or potential conclusion of a formal logical statement. At the same time, this does not mean that this assertion based on analogy is necessarily unfounded or irrelevant. As Pesce (1999) states “meaning cannot be counted”, i.e. translated into an unambiguous language that the computer can process.

Third, the demands placed on this type of reasoning are ambitious – “to solve problems in domains that ordinarily require human logical reasoning” (cf. Ludwig 2010) – but the machines” ability to facilitate such inference on a large scale (and perhaps also to a humanist’s standards) is often limited and contentious (cf. Zöllner-Weber 2009, cf. McCarty 2005). In this context, the modelling of a domain in such a way so that inference machines can eventually create valuable knowledge from it is an activity that is dependent on a large investment of human reasoning in the first place. The information to be “discovered” has to be preconfigured in the knowledge representation. Perhaps the combining of the dataset with the ontology can lead to the computer being able to “infer” that “Flipper isA Mammal”, but only because humans “know” this in the first place. What does a humanist gain by intensively modelling a domain so that the computer can discover what she already knows?

6.1.2 Reasoning in the Humanities

Providing an extensive analysis of humanist reasoning practices would go far beyond the scope of the deliverable and would be an elaborate scholarly endeavour in and of itself. Our purpose here is instead to point to other ways of thinking about the term “reasoning” for the Semantic Web that are more familiar to the humanist, so as to shift the weight of the discussion towards a position that includes the value of humanist input about thinking about data created in the Web environment, especially in environments using Linked Data. For this reason we will first provide a much broader definition of reasoning than the one given above. This will hopefully help to highlight the potential of humanist ways of looking at the data, which we will then explore in our use cases in the next section.

The Oxford Handbook of Thinking and Reasoning generically defines the term “reasoning” as follows: “Reasoning, which has a long tradition that springs from philosophy and logic, places emphasis on the process of drawing inferences (conclusions) from some initial information (premises)” (Holyoak and Morrison 2012: 2). According to the authors, reasoning is intrinsically related to but not necessarily synonymous with the act of thinking in general and closely tied to many other mental activities such as judgment, decision making, creative thinking and problem solving (cf. Holyoak and Morrison 2012: 2). Although the rigorous confines of formal logic are used in some reasoning practices such as those of (analytic) philosophy and mathematics, there are other practices or “styles” of reasoning that place much less emphasis on this (cf. McCarty 2006, cf. Crombie 1994 and Hacking 1985). Indeed, logic is sometimes considered an attempt to provide a normative model for the reasoning process (Holyoak and Morrison 2012: 4-5) or a “grammar of thought” (Zoglauer 2008: 9), but it is not to be confused with all of the complex cognitive processes and scientific practices involved in reasoning itself.

Crombie (1994: 155) explicitly makes a distinction between logic and reasoning. He states: "First, I observe that by reasoning I don't mean logic. I mean the very opposite, for logic is the preservation of truth, while a style of reasoning is what brings in the possibility of truth or falsehood." While the exact wording of this distinction may be contentious, we find the general tenor of the statement to be of value, that logic and reasoning are not synonymous: logic is a tool that can be used to make sure that arguments are sound, but reasoning involves the entire process of coming to conclusions and is dependent on different scientific cultures.[112] McCarty (2008: 12), referencing Crombie (1994), talks about different cultural practices which have evolved to help humans come to conclusions as "styles" of reasoning. In this respect he provides us with the following list:

"The simple method of postulation exemplified by the Greek mathematical sciences; The deployment of experiment both to control postulation and to explore by observation and measurement; Hypothetical construction of analogical models; Ordering of variety by comparison and taxonomy; Statistical analysis of regularities of populations, and the calculus of probabilities; The historical derivation of genetic development."

Keeping this in mind, Holyoak and Morrison (2012) discuss a number of different aspects that need to be considered when talking about "reasoning" in a wider sense. These include not only the different scientific approaches that have normalised reasoning processes, but also the different methods of coming to conclusions (inductive, deductive, abductive), the intricacies of judgement and decision making, the impacts of language and culture on reasoning, and different modes and practices of thinking.

This brings us to our plea for widening the understanding of reasoning in the digital humanities, especially related to the Semantic Web, to include other styles of reasoning with the data than the purely computer-oriented ones mentioned in the previous section. Since human reasoning is ultimately at the basis of any reasoning programme, understanding what reasoning practices humanists may engage in with the available data can also ultimately help the future implementation of reasoning in the computer science sense. It is also important to discover the ways humanists are able to reason (come to conclusions) using the simple tools and functionalities immediately available to them.

For our purposes then, the definition of Semantic Web "reasoning" will be extended to include any process of interaction with (Linked) Data and the resulting graph that leads to the discovery of new information and the potential creation of new triples. Note that this definition does not restrict reasoning to drawing inferences, but still has a focus on coming to conclusions. This interaction can be either driven by human or computer interaction. In our case we will concentrate on human interaction with the graph, specifically on the technology of the faceted browser. This will be discussed in more detail in the following section.

6.1.3 Interactive Reasoning

Although we have made the definition of Semantic Web reasoning for this deliverable very broad, our vision for exploring what we call "interactive reasoning" for the Semantic Web is narrower in scope. Here we will concentrate on how humans can use the simple but effective

[112] To be sure, there is much overlap between styles, modes or practices of reasoning and it should not be implied here that these types of reasoning are somehow less rigorous than the types of reasoning specific to mathematics, philosophy and the natural sciences (cf. Holyoak and Morrison 2012: 11). As Holyoak and Morrison (2012: 3) point out, Thomas Hobbes, in designating reasoning as form of "reckoning", equated it with "computations," as in arithmetic calculations". This notion of thinking and reasoning has persisted within the sciences (Holyoak Morrison 2012: 3ff.) and is not unfounded. Indeed this implies a type of computation that is at the heart of every reasoning practice.

tools that exist for exploring and exploiting the Linked Data graph in order to come to new conclusions (i.e. create new triples) about the data. Through our use cases we have discovered the potential for studying "interactive reasoning" for scholarship on the basis of Linked Data in the humanities in this regard on at least three levels. Reasoning as a human cognitive activity is involved in

1. selecting, modifying, or creating (annotation) vocabularies for particular data sets and research interests;
2. applying the (annotation) vocabulary by annotating resources;
3. exploring and assessing the data by visualising and querying the graph that has been established through the creation of an (annotation) vocabulary and the annotation of a data set.

This third step can not only be carried out by a reasoner, but also by a human using other tools such as faceted browsers. Our case studies will reflect these areas of humanist "interactive reasoning". The next section will discuss reasoning with faceted browsers in general.

Using, among other things, the scholarly research platform based on the Pundit family of tools, DM2E has been conducting research into current and potential scholarly practices with Linked Data (cf. chapter 2). Central to the current task on "reasoning" has been to further explore the useful but still limited capabilities of the Ask faceted browser by having data created by DM2E scholars made searchable using additional faceted browsers implemented for their specific purposes. Scholars could then use the results to make inferences about the data created, potentially coming to new conclusions, discovering new information and creating new links.

An overview of the functionalities of the Pundit family of tools used for creating the datasets in the case studies can be found in other DM2E publications.[113] Here, we will briefly discuss Ask, as it includes the potential of the faceted browser. Ask provides a domain independent view on annotations (Cf. D3.3) created in Pundit. It is used for managing personal notebooks containing annotations, viewing notebook contents, and providing a basis for simple vertical visualisations. The notebooks faceted browser, which allows for any number of notebooks to be searched dynamically, is a very powerful feature for analysing a corpus of annotations. But since Ask is domain independent, it provides only a generic way of exploring the whole graph and the facets are limited to the particular instance data of the subjects and objects, the properties, and class types.

A faceted browser can however, be tailored to a specific dataset. In general, it is an application that allows the user to access data using different filters. The data can then be combined and recombined in different ways depending on the chosen filters, providing novel ways of looking at the data. Faceted browsers have a distinct advantage for scholars from humanities" domains wanting to explore and query information stored as Linked Data. For one, they provide an immediately accessible but structured visualisation of the specific subjects, objects and predicates in the data set. Browsing, in contrast, an RDF/XML document per se is not the easiest way to make sense of the data. In addition, if a faceted browser is provided, other scholars who are perhaps less technically inclined can obtain this overview of the contents and structure of the data set without needing any knowledge of programming languages or query languages such as SPARQL. They can therefore solely concentrate on comparing their understanding of the domain with the one represented by the data set. In short, a faceted browser facilitates reasoning by permitting the scholar to relatively quickly identify, in a given dataset, the data and metadata most relevant for the

[113] Cf., for example, Grassi et al. (2013).

pursuit of a specific research question by iterative processes of selecting and deselecting given facets. The facets offered at each time for selection will always be a result from exactly the previous selections and thus guide the scholar through and document the research path. Such reasoning by faceted browsing is to us an example of a kind of reasoning that humanists can perform with benefit in the context of the Semantic Web and the Digital Humanities, even though it is not fully automated reasoning, and thus not Semantic Web reasoning in the standard sense.

For our experiments, we had scholars load their data into faceted browsers tailored to their specific research data and domain. In the next section we will discuss reasoning with Linked Data based on two case studies, including ones utilising faceted browsers.

6.2 Reasoning Use Cases

For our research into “interactive reasoning” with Linked Data we analysed case studies (Wittgenstein Archives at the University of Bergen, Georg Eckert Institute) from two separate digital humanities domains (philosophy, history), for which we provided self-documentation forms and conducted expert interviews. This section will discuss the two case studies in detail and then explore the reasoning scenarios they entailed.

6.2.1 GEI Case Study

For the first case study to be examined in this deliverable, DM2E worked with a scholar at the Georg-Eckert-Institute (GEI), a DM2E associate partner, from the field of educational history. An experiment with Pundit and its components was set up for this purpose. The experiment took place between September and November 2014. The scholar was given an introduction to important Linked Data concepts and methods and asked to define a relevant use case from his particular interest area in the field of history. The participant was asked to determine the nature of the semantic annotations to be established and the sources to be used based on his usual research methods. Then, the participant created about 250 semantic annotations on relevant historical digital materials, which resulted in small but meaningful graphs. The process can be described as a basic and intuitive translation of common research methods from the humanist field of history into the Linked Data paradigm for the purpose of answering research questions specific to a humanist scholar’s area of interest and expertise. For this reason it served, among other things, as a use case for analysing the way in which humanists might want to “reason” with Linked Data. For the purposes of the “reasoning” aspect of the experiment, the scholar was asked to explore the graph using a faceted browser and given a self-documentation form to complete (Addendum), which had been previously prepared by the members of DM2E in the context of Task 3.4.

Method

The source documents for the experiments were taken from the digital library of GEI and consisted of different types (e.g., protestant or catholic) of historical school books from Germany published from ca. 1850 to 1900.[114] The (RDF) metadata of these sources have

[114] The corpus of GEI-Digital was created by the Georg-Eckert-Institute for International Textbook Research Member of the Leibniz Association. It was created to be used by different scientific communities (e.g. historians, education researchers). The aim is to permit an easy access to this kind of source material; to capture them in form and content and to make full-text versions available to a wider, international user group. The parallel aim is long-term sustainability of the books themselves. It contains images, bibliographical data, context, full text, a pdf version, esp, mets/mods.

also been provided to Europeana via DM2E[115]. The particular research method was hermeneutic and involved closely reading the historical sources and identifying various topoi (salient terms), and their connotation (positive or negative) and presentation in the different school books. Specifically, the experiment focused on the questions: "Which topoi appear in which textbooks?", "How are they connotated?" and "In which context have they been set?" Topoi were annotated according to several criteria including the nature of the connotation. The desired result of the annotations was to be able to compare the topoi in different texts over time, assessing, for example, which topoi can be found in which documents and how the connotation and frequency of certain topoi change over time.

The scholar is part of a research group that works on and with textbooks and juvenile literature of the nineteenth century. Because textbooks are semi-official documents that were read by wider parts of the Germans during their formative years, his group tries to find the representations of the world and the nation and the description of historical processes that were offered by the state to its future citizens. So, they search for representations of the nation and the globalised world. Also, they look for representations of change, crisis, religious conflict, social change and similar events.

The experiment involved a three-step process consisting of a 1) source critique (documented with Pundit), 2) content analysis (documented with Pundit) and 3) exploration of the data with a faceted browser (Ask). The first two steps are based in the hermeneutical method and translated to the Linked Data paradigm. A detailed analysis of the results will be available in the form of a paper in the future. The last step involves using the automatic visualisation and combination functionalities of the faceted browser to evaluate the results and therefore the reasoning process involved.

The source critique consisted of skim reading the texts and establishing an initial contextualisation. This involved selecting sources from the digital library of GEI, reading the sources based on experience and hypotheses, identifying noticeable entities in the text and collecting potentially interesting texts in a notebook as a sample by creating annotations with Pundit. The annotations used in the first stage of the source critique identified facts such as

- place of publication
- year of publication
- edition
- author
- publisher
- religious attribution of the text (e.g. protestant, catholic, neutral)
- school type (e.g. girl's school, boy's school, gymnasium, teacher's seminar)
- regional attribution (e.g. for schools in Baden)
- discipline (e.g. history, geography.)

These basic attributes were the building blocks for establishing context by connecting them with further facts about the corresponding historical environment, which included

[115] http://www.europeana.eu/portal/search.html?query=*%3A*&rows=24&qf=PROVIDER%3ADM2E&qf=DATA_PROVIDER%3A%22Wittgenstein+Archives+at+the+University+of+Bergen+%28WAB%29%22&qt=false

- relevant historical events, such as the foundation of the German Reich in 1871
- historical periods: ~1850-1870, 1871-1884, 1885-1900, 1901-1914, 1914-1918
- names of smaller sub-periods (German: “Querperioden”), such as the period of the Socialist Laws (German: “Sozialistengesetze”) from 1878 to 1890.

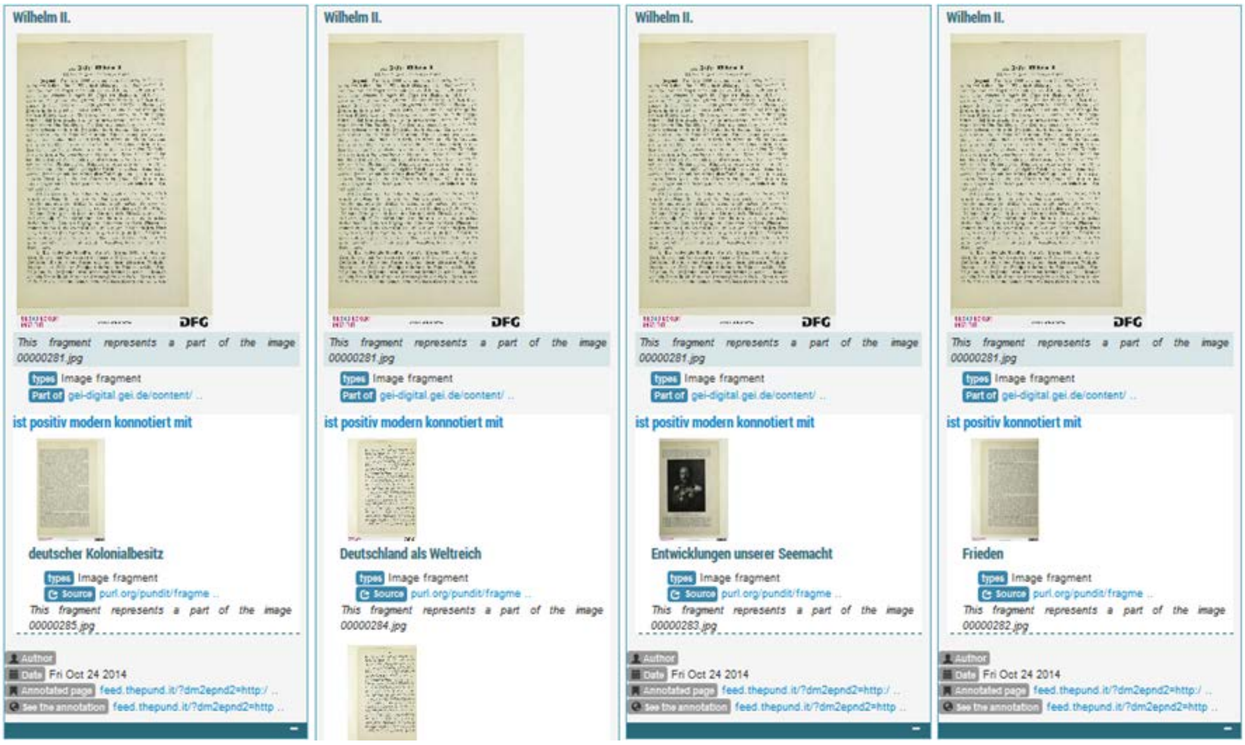
In addition to annotating attributes of elements within the documents and contemporary events corresponding to the time period they were in use, the scholar was also interested in analysing the content according to the implicit values and opinions contained within them. For this purpose, the second step involved content analysis, which meant looking for evaluative concepts. First, pertinent “topoi" and "connotations" were identified and annotated using RDF triples with the form

- :x :is_a :topoi
- :x :is_used_as :connotation (e.g. “anti-secular”, “anti-religious”)
- :x :is_used_with :positive_connotation, negative_connotation

Second, paraphrases of the topoi were identified and annotated:

- :y is_a :paraphrase
- :y :refers_to :topoi

This process in terms of Linked Data can be described as the creation of a small-scale vocabulary for the purpose of documenting evaluative statements found in historical texts. The results of this experiment can be currently viewed in the faceted browser.[116] One example of the triples created can be illustrated with following screenshot (figure 18) taken from the scholar’s Pundit Notebook entitled “Welt der Kinder”:



Wilhelm II.

This fragment represents a part of the image 00000281.jpg

types Image fragment
Part of gei-digital.gei.de/content/ ...

ist positiv modern konnotiert mit

deutscher Kolonialbesitz

types Image fragment
Source purl.org/pundit/fragme ...
This fragment represents a part of the image 00000285.jpg

Author
Date Fri Oct 24 2014
Annotated page feed.thepund.it/?dm2epnd2=http:/ ...
See the annotation feed.thepund.it/?dm2epnd2=http ...

Wilhelm II.

This fragment represents a part of the image 00000281.jpg

types Image fragment
Part of gei-digital.gei.de/content/ ...

ist positiv modern konnotiert mit

Deutschland als Weltreich

types Image fragment
Source purl.org/pundit/fragme ...
This fragment represents a part of the image 00000284.jpg

Wilhelm II.

This fragment represents a part of the image 00000281.jpg

types Image fragment
Part of gei-digital.gei.de/content/ ...

ist positiv modern konnotiert mit

Entwicklungen unserer Seemacht

types Image fragment
Source purl.org/pundit/fragme ...
This fragment represents a part of the image 00000283.jpg

Author
Date Fri Oct 24 2014
Annotated page feed.thepund.it/?dm2epnd2=http:/ ...
See the annotation feed.thepund.it/?dm2epnd2=http ...

Wilhelm II.

This fragment represents a part of the image 00000281.jpg

types Image fragment
Part of gei-digital.gei.de/content/ ...

ist positiv konnotiert mit

Frieden

types Image fragment
Source purl.org/pundit/fragme ...
This fragment represents a part of the image 00000282.jpg

Author
Date Fri Oct 24 2014
Annotated page feed.thepund.it/?dm2epnd2=http:/ ...
See the annotation feed.thepund.it/?dm2epnd2=http ...

Figure 16. Triple-display, “Welt der Kinder” notebook in Pundit.

[116] http://demo-search.thepund.it/

The four textual examples all come from the same textbook: "German History from the Migration Period to the Present" (German: "Deutsche Geschichte von der Völkerwanderung bis zur Gegenwart") by Ludwig Kahnmeyer, Adolf Dilcher and Hermann Schulze, which was published 1913. They deal with the "topos" of Wilhelm II, the German Emperor who ruled the German Empire and the Kingdom of Prussia from 1888 to 1918. The textbook therefore was published during his reign. The historian created several annotations for this topos (triple subject), which occurs on several different pages of the textbook (pp. 281-285). It is positively connotated (triple predicate/relation) with the concepts (triple objects) "development of our naval power" (German: "Entwicklungen unserer Seemacht"), "Germany as a world power" (German: "Deutschland als Weltreich"), "the possession of German colonies" (German: "deutscher Kolonialbesitz") and "peace" (German: "Frieden").

This small section of the work done in this experiment shows an example of the results of the researcher's methods. For one, we see several aspects of the official German state propaganda of the time, which here are expressed as pride in the military and an attribution of state policies to the figure of the Emperor. In the larger context of state propaganda in general, we can clearly see the often overlooked contradictory nature of ideology: a political figure can be positively connotated with militaristic concepts such as "naval power" or "colonialism" and, at the same time, with "peace".

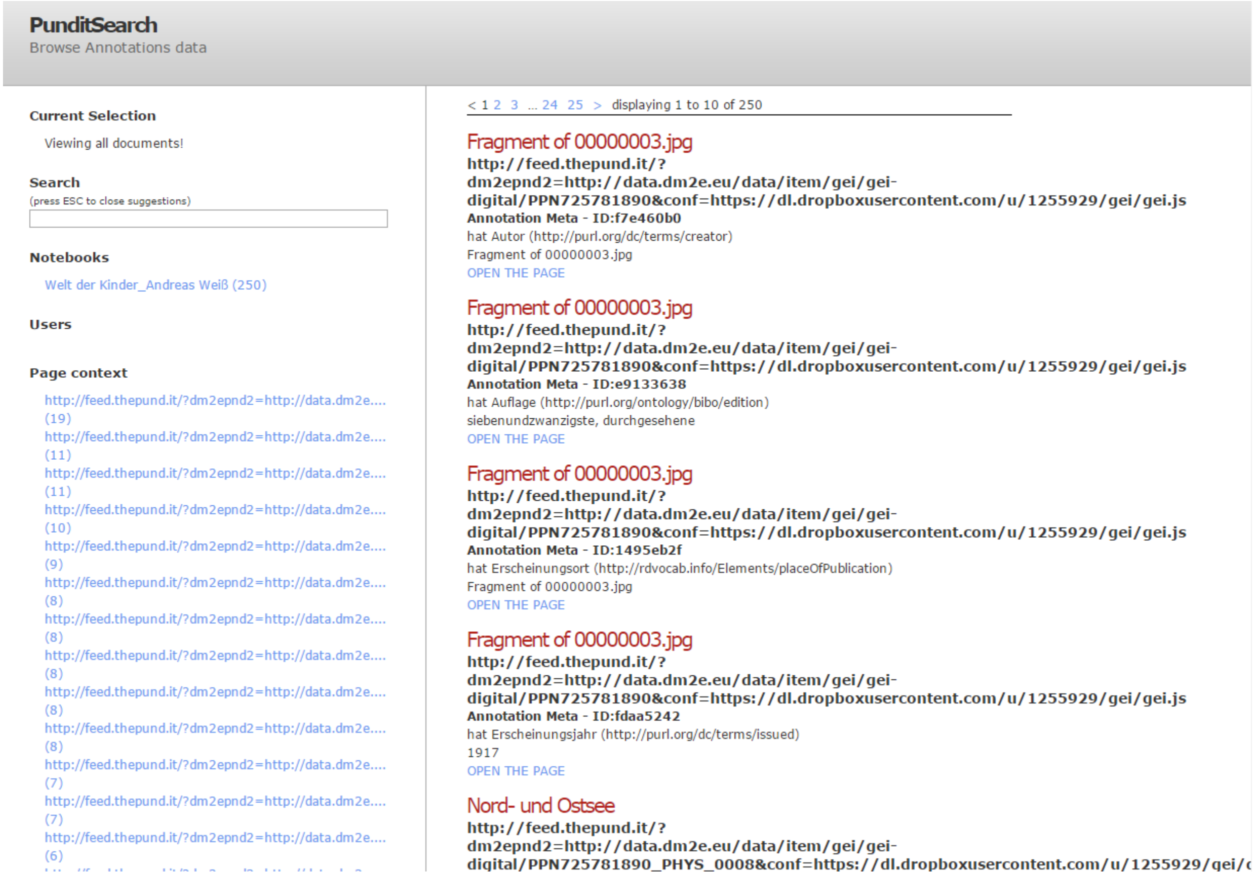
It is evident from the example annotations concerning Wilhelm II given above that the first two steps of the experiments based on the historian's hermeneutic method involving the close reading of a source were able to deliver telling results about the research object. For these two steps, the scholar did not necessarily need the aid of any computer technology at all. However, the simultaneous utilisation of simple Linked Data tools (creating semantic annotations with Pundit) provided him with the basis for new ways of storing and displaying his data that can lead to novel ways of looking at, working with and reasoning on the results obtained.[117] This was explored in the third step of the experiment: exploring the data with Ask.

Although the scholar did not create an elaborate ontology to represent his particular domain, he did establish a small vocabulary to explicitly document both historical facts and statements about meaning concerning the content of his research object. The evidence behind his conclusions as well as the conclusions themselves, captured in the form of triple statements, could therefore be automatically recalled using Ask and a simple faceted browser, PunditSearch,[118] built specifically for this purpose.

In the third step, the scholar was asked to use the faceted browser to query the data with regards to a specific research question about the source and the Linked Data created in the first two steps. The faceted browser allows to incrementally filter triples based on the instances found in the subject, predicate, and object position of triples, and according to classes. The browser adjusts the display of triples matching the selected facets. Below is a screenshot of the PunditSearch faceted browser.

[117] The scholar did not involve heavier Semantic Web Reasoning methods such as the creation of an ontology, computer algorithms or formal logic to come to his conclusions or achieve his results.

[118] http://demo-search.thepund.it/



PunditSearch
Browse Annotations data

Current Selection

Viewing all documents!

Search
(press ESC to close suggestions)

Notebooks

Welt der Kinder_Andreas Weiß (250)

Users

Page context

http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (19)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (11)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (11)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (10)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (9)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (8)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (8)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (8)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (8)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (8)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (7)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (7)
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.... (6)

< 1 2 3 ... 24 25 > displaying 1 to 10 of 250

Fragment of 00000003.jpg
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.eu/data/item/gei/gei-digital/PPN725781890&conf=https://dl.dropboxusercontent.com/u/1255929/gei/gei.js
Annotation Meta - ID:f7e460b0
hat Autor (http://purl.org/dc/terms/creator)
Fragment of 00000003.jpg
OPEN THE PAGE

Fragment of 00000003.jpg
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.eu/data/item/gei/gei-digital/PPN725781890&conf=https://dl.dropboxusercontent.com/u/1255929/gei/gei.js
Annotation Meta - ID:e9133638
hat Auflage (http://purl.org/ontology/bibo/edition)
siebenundzwanzigste, durchgesehene
OPEN THE PAGE

Fragment of 00000003.jpg
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.eu/data/item/gei/gei-digital/PPN725781890&conf=https://dl.dropboxusercontent.com/u/1255929/gei/gei.js
Annotation Meta - ID:1495eb2f
hat Erscheinungsort (http://rdvocab.info/Elements/placeOfPublication)
Fragment of 00000003.jpg
OPEN THE PAGE

Fragment of 00000003.jpg
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.eu/data/item/gei/gei-digital/PPN725781890&conf=https://dl.dropboxusercontent.com/u/1255929/gei/gei.js
Annotation Meta - ID:fdaa5242
hat Erscheinungsjahr (http://purl.org/dc/terms/issued)
1917
OPEN THE PAGE

Nord- und Ostsee
http://feed.thepund.it/?dm2epnd2=http://data.dm2e.eu/data/item/gei/gei-digital/PPN725781890_PHYS_0008&conf=https://dl.dropboxusercontent.com/u/1255929/gei/c

Figure 17. Screenshot of the PunditSearch faceted browser displaying the GEI dataset.

At the same time, the scholar was given a self-documentation form in which he was asked to describe and reflect on the process. The following is a summary and analysis of the documentation.

Our scholar decided first to obtain an overview of the object labels, as he found oversight of the triples created to be difficult. After browsing the available object labels, he chose to take a closer look at the most numerous (with 5 instances), the “German Empire”. The list of related (triple) subjects returned by the browser surprised him, as the German Empire was only connoted with what he felt were “internal topics” (national as opposed to international). He had expected the object “German Empire”, however, to also be compared with “external” topics such as “France” and other “Empires”.

This first look at the object label group raised questions for our historian, which could be considered part of a reasoning process that can lead to the creation of new triples. As the scholar looks at the subjects returned as a result of the faceting browsing, he compares them to his expectations, which are based on real-world knowledge. As a consequence, he discovers a similarity in the results that occurs to him because of the absence of what he expected to be in the results. This is quite a dynamic, intuitive and partially serendipitous reasoning process. Before looking at the group, the scholar might not yet have known what his expectations were, i.e. that the category of “national” versus “international” is relevant when describing the way in which textbooks talked about the “German Empire”. He now can make this implicit information explicit by creating a new triple expressing this (a product of reasoning as defined in this deliverable). In addition, he already has a result and can use his knowledge to further determine what it means: “One way the topos of the “German Empire” is constructed/talked about in the school books is by listing positive attributes related to national subjects”.

The exploration (faceted browsing) of the object label “German Empire” returned only positively connotated results. This unexpected pattern led the scholar to take a look at all the topoi that are connotated negatively (43) and all that are connotated positively (63). This additional step occurred out of a desire to gather more data to understand the results. Here, an unexpected pattern causes the scholar to look for other significant patterns, relationships and groupings in the dataset of positively and the negatively connotated topoi.

In the process he discovers that one of the negatively connotated subjects is France. Being a European country, this subject can be considered a similar but distinct counterpart to the object label “German Empire”. The scholar’s choice to look at France therefore has a basis in analogy. The scholar has a look at the topoi (subjects) associated with France, discovering that the negative connotation has to do with France’s naval power and its open borders. Combining this with his real-world knowledge, he comes to the conclusion that this contrast is an expression of “military and political rivalry.” He compares this new information to his previous results about the “German Empire” and also comes to a conclusion, that an antagonism is created in the textbooks between the “German Empire as a supporter of peace and France as an aggressive and potentially dangerous neighbour”.

Once again the results make the scholar want to explore more data. With each conclusion about a certain subject he extends the search to other analogous elements of the dataset in an iterative process. The scholar therefore resets the facets to search for what connotations have been made about other neighbours of the “German Empire”. The results lead the scholar to believe that most countries are constructed as potential rivals of Germany.

The scholar’s final conclusion from this search is that “the German nation is represented as a modern peaceful one that exists unfortunately in a dangerous environment. And crises and potentially dangerous changes loom everywhere!”

Results

In the scope of this deliverable it is impossible to make sweeping statements about what types of Semantic Web reasoning all humanists want to see employed with Linked Data, but we can use this case study to make empirical observations about the practice or style of reasoning found at the intersection between traditional humanistic research practices and those (that will be) made possible through the use of Linked Data tools.

In order to be able to determine the style of reasoning expressed in this use case, three relevant aspects of the experiment should be discussed. First, one of the difficulties of Semantic Web reasoning mentioned previously in the deliverable was that the research objects and research questions of humanities scholars are qualitatively different than those of scholars of mathematics and life sciences. For this reason, the first aspect considered will be the scholar’s chosen research object and research question. The second aspect will discuss the underlying research method used in the creation of the Linked Data, which can be described as being at the intersection of humanist and Linked Data methods. Lastly, we will discuss how our scholar used the faceted browser tool to come to conclusions about the RDF data set he created and how he assesses the method.

A look at the object of research and research question in this case study supports the idea that humanists are interested in meaning. The scholar was interested in historical facts, but more importantly he wanted to study the construction and expression of opinions, values, worldviews, and biases in the historical school books. Historical facts were important for providing the context of the value “statements”, but addressing their meaning to the authors and potential influence on the recipients of the works was the most important aspect. As a result, the small vocabulary created by the scholar was primarily tailored to the documentation of attitudes. Although statements created are useful, they are not necessarily axiomatic.

The underlying research method and therefore perhaps also reasoning style of the scholar was by self-admission hermeneutic. It involved close reading and textual interpretation based on real-world knowledge, uncovering subtext, and searching for meaning. This was then combined with the Linked Data methods by having the scholar explicitly state the results of this close reading in his annotation. Meaning is therefore interpreted in this process and stated using triples.

The scholar's self-documentation of his interaction with the graph using the faceted browser gives us some insight into how humanists can use their methods to interact with and ultimately reason with Linked Data. More specifically, we are shown how they can draw conclusions from the data using the faceted browser. Of course this type of reasoning is limited to and contingent upon the specific technological implementation. It is, however, instructive in uncovering a complex and dynamic humanistic reasoning process using Linked Data.

In general, we have noticed a main procedure of reasoning with Linked Data using facets that is iteratively repeated: the scholar analyses the results (of applying certain facets to the data) by comparing them with his real-world/scientifically acquired knowledge in different ways and creating hypotheses about them based on this. With the faceted browser he can then re-shuffle facets to find how other combinations undermine or support hypotheses and to look for answers to new questions that arise. This reasoning procedure is conducted on the basis of statements the scholar created himself, i.e. the annotation vocabulary and instance data, which allows him to better comprehend the context of what he sees.

Real-world knowledge means keeping in mind that results shown in the faceted browser are reflective of certain assumptions and biases explicitly and implicitly addressed within the texts that can be determined by considering their context. Context is, of course, more than just referencing the name and vita of the author(s) and sponsor(s) or the dates of the time period it was written, but a deep understanding of what hidden agendas the authors had, what values they were trying to perpetuate, and what this all means for people living at the time of publication as well as today. The significance or relevance of certain aspects of context shifts according to the question asked, who asks it, for what purpose and the results given.

When looking at the results, the scholar observes, for example,

- (relevant) patterns in the data
- a pattern/information that supports expectations
- a pattern/information that contradicts expectations arrived from judging the context
- unexpected information
- salient information
- anomalous information
- absurd information
- analogous resources/information
- antagonistic resources/information

and he compares and contrasts this with real-knowledge and experience.

The scholar's own criticism of the resulting reasoning process is valuable for assessing the value and potential uptake of this kind of "interactive reasoning" with Linked Data.

On a positive note, the scholar cited several positive aspects of this approach. He felt it aided thoroughness through forcing the repetition of statements. It aided his ability to quickly compare the data created from the close reading of several sources. A hypothesis could be instantly tested and results automatically reproduced. The recall of relevant information was therefore much faster.

6.2.2 Wittgenstein Ontology Case Study

The second use case provides a further perspective on the reasoning topic. In contrast to the use case with the GEI, the researchers at the Bergen Wittgenstein Archives (WAB) had previously created an ontology (using RDF-triples) for use with the data in its archive, the so-called Wittgenstein Ontology (WO). This ontology was developed by digital humanists with knowledge of Semantic Web technologies. The use case focused on the characteristics of the ontology and on the interaction of two scholars with the ontology using the faceted browser.

Method

WAB is a partner of DM2E, providing a digitised edition of Wittgenstein's Nachlass, which is produced from WAB's machine-readable version of Wittgenstein's Nachlass.[119] The Wittgenstein Nachlass amounts in total to ca. 20,000 pages, while the Wittgenstein Source corpus on wittgensteinsource.org includes a 5,000 page selection from this larger corpus. Wittgenstein Source was created in the framework of the Discovery project[120] by WAB for Open Access Wittgenstein research. It contains English and German manuscripts and typescripts from Wittgenstein's Nachlass in facsimiles and as diplomatic and normalised text editions. It also contains metadata and short descriptions of these items.

The WO is linked with Wittgenstein Source, partly through Pundit. The Linked Data representation of the ontology was created by WAB for both internal and external use. Internal use includes checking of metadata comprehensiveness and consistency, external use (by researchers) includes searching and browsing of metadata. The ontology was intended primarily to assist Wittgenstein research. It includes classes for primary and secondary sources, concepts and persons. The lowest subclass of a Wittgenstein primary source is the Bemerkung; the Bemerkung denotes, roughly speaking, a single Wittgensteinian remark (German: "Bemerkung"). Instances of the different classes are interlinked with each other via properties / predicates.[121]

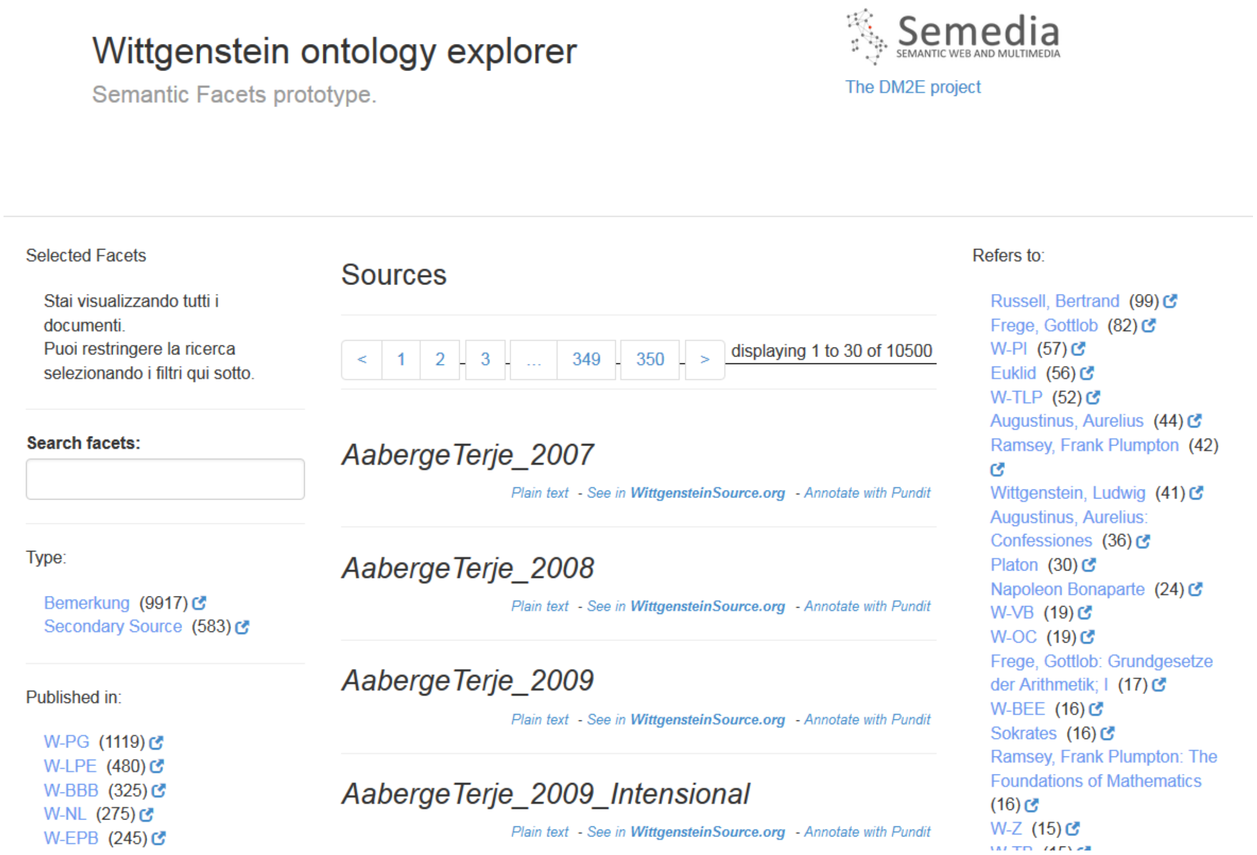
For querying the data set of the WO, Net7 worked with WAB to create the Wittgenstein Ontology Explorer,[122] which is a semantic facets browser using the open source software Ajax-Solr.[123] Users can choose from the following facets (subjects, objects and predicates in the ontology) or search for terms in the facets using a search bar:

- Type (source category - primary or secondary)

[119] Ref. to the Bergen Electronic Edition.
[120] http://wab.uib.no/wab_discovery.page and http://www.discovery-project.eu/home.html.
[121] Cf. Pichler and Zöllner-Weber (2012).
[122] http://141.20.126.236/dm2e/ajax-solr/examples/wab/
[123] https://github.com/evolvingweb/ajax-solr

- Published in (work)
- Part of (manuscript, typescript ...)
- Date (of remark)
- Source (secondary or primary)
- Refers to (person)
- Discusses (topic)
- Other version (of remark)

Below is a screenshot (figure 20) of the faceted browser, in which no facets value have been selected.



Wittgenstein ontology explorer

Semantic Facets prototype.

Selected Facets

Stai visualizzando tutti i documenti.
Puoi restringere la ricerca selezionando i filtri qui sotto.

Search facets:

Type:

Bemerkung (9917)
Secondary Source (583)

Published in:

W-PG (1119)
W-LPE (480)
W-BBB (325)
W-NL (275)
W-EPB (245)

Sources

< 1 2 3 ... 349 350 > displaying 1 to 30 of 10500

AabergeTerje_2007

Plain text - See in WittgensteinSource.org - Annotate with Pundit

AabergeTerje_2008

Plain text - See in WittgensteinSource.org - Annotate with Pundit

AabergeTerje_2009

Plain text - See in WittgensteinSource.org - Annotate with Pundit

AabergeTerje_2009_Intensional

Plain text - See in WittgensteinSource.org - Annotate with Pundit

Refers to:

Russell, Bertrand (99)
Frege, Gottlob (82)
W-PI (57)
Euklid (56)
W-TLP (52)
Augustinus, Aurelius (44)
Ramsey, Frank Plumpton (42)
Wittgenstein, Ludwig (41)
Augustinus, Aurelius: Confessiones (36)
Platon (30)
Napoleon Bonaparte (24)
W-VB (19)
W-OC (19)
Frege, Gottlob: Grundgesetze der Arithmetik; I (17)
W-BEE (16)
Sokrates (16)
Ramsey, Frank Plumpton: The Foundations of Mathematics (16)
W-Z (15)
W-TP (15)

Figure 18. The faceted Wittgenstein Ontology Explorer.

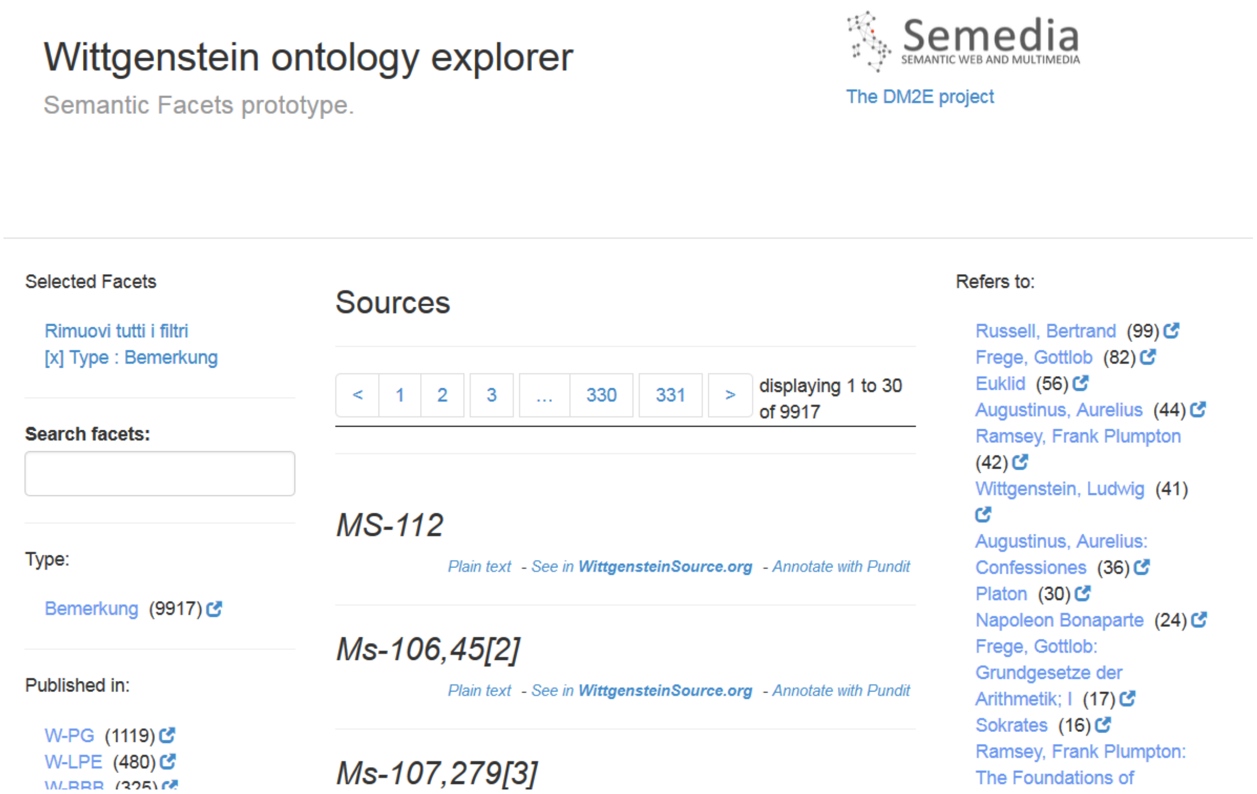
For this use case, we asked two Wittgenstein scholars to use the faceted browser to answer a particular research question about Wittgenstein’s oeuvre; they were allowed to choose the question and then given the Self-Documentation worksheet to capture the research/reasoning process. The digitised form of the Nachlass on Wittgenstein Source constituted the basis for both investigations. In following, the results of both of these experiments documented by the scholars will be described. In addition, the working group carried out open expert interviews with these two scholars about their results. These will also be included, as they were very telling for the purposes of the reasoning experiment.

Scholar One: Wittgenstein’s concept of philosophy

The first scholar who completed this experiment (February-April 2014) was already familiar with the ontology, as he is one of the researchers responsible for its creation. His chosen research question and method were partly a simulation of how he imagined another Wittgenstein scholar might use the ontology for its intended purpose. In general, the research question behind the experiment involved imagining how a scholar (perhaps a student) could be assisted by the ontology explorer to come to an understanding of Wittgenstein’s conception of philosophy. The associated research method was to explore, analyse and compare primary and secondary sources on the subject by among others exploring key concepts in the ontology. The tools the researcher had at his disposal were the digitised version of Wittgenstein’s Nachlass contained on Wittgenstein Source and a faceted browser for exploring the WO, which is linked to the Nachlass.

Finding key texts in Wittgenstein‘s Nachlass

The first step in the experiment lead the scholar to try to identify key texts relevant for the question what Wittgenstein thinks about philosophy. To do this, he chose the facet “Type” and selected the value “Bemerkung” in the faceted browser (figure 21).



Wittgenstein ontology explorer

Semantic Facets prototype.

Semedia SEMANTIC WEB AND MULTIMEDIA

The DM2E project

Selected Facets

Rimuovi tutti i filtri
[x] Type : Bemerkung

Search facets:

Type:

Bemerkung (9917)

Published in:

W-PG (1119)
W-LPE (480)
W-BBB (325)

Sources

< 1 2 3 ... 330 331 > displaying 1 to 30 of 9917

MS-112
Plain text - See in WittgensteinSource.org - Annotate with Pundit

Ms-106,45[2]
Plain text - See in WittgensteinSource.org - Annotate with Pundit

Ms-107,279[3]

Refers to:

Russell, Bertrand (99)
Frege, Gottlob (82)
Euklid (56)
Augustinus, Aurelius (44)
Ramsey, Frank Plumpton (42)
Wittgenstein, Ludwig (41)
Augustinus, Aurelius: Confessiones (36)
Platon (30)
Napoleon Bonaparte (24)
Frege, Gottlob: Grundgesetze der Arithmetik; I (17)
Sokrates (16)
Ramsey, Frank Plumpton: The Foundations of

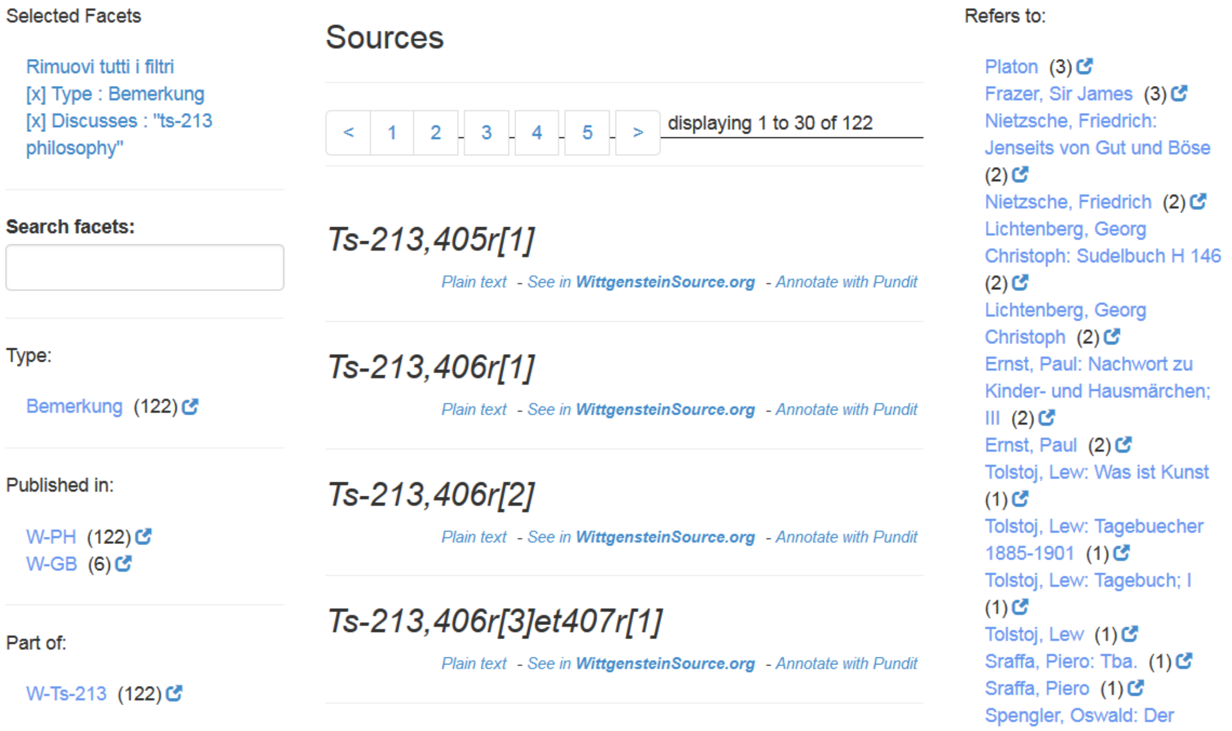
Figure 19. Applying the facet “Type: Bemerkung”.

Faceted browsers act as a filter on the data. Choosing a facet value means restricting the search results to only those resources that are associated with that value. By selecting the value “Bemerkung” under the facet “Type”, the scholar has eliminated all secondary sources mentioned in the ontology – there are only two “types” in the ontology, see below – from the current view of the faceted browser. “Bemerkung” is an object of the predicate “:hasType”. The subject (X) of this triple is the resource representing a text section.

- :X :hasType :Bemerkung

- :X :hasType :Secondary_Source

After this first filtering, the scholar uses the search bar to look for the word “philosophy”, and finds out that this word matches one value under the facet “Discusses”. In doing so, the scholar discovers a Bemerkung in Wittgenstein’s Nachlass that WAB’s Wittgenstein ontology indicates as containing a “discussion” of “philosophy”. This passage is in TS-213.



Selected Facets

Rimuovi tutti i filtri
[x] Type : Bemerkung
[x] Discusses : "ts-213 philosophy"

Search facets:

Type:

Bemerkung (122)

Published in:

W-PH (122)
W-GB (6)

Part of:

W-Ts-213 (122)

Sources

< 1 2 3 4 5 > displaying 1 to 30 of 122

Ts-213,405r[1]

Plain text - See in WittgensteinSource.org - Annotate with Pundit

Ts-213,406r[1]

Plain text - See in WittgensteinSource.org - Annotate with Pundit

Ts-213,406r[2]

Plain text - See in WittgensteinSource.org - Annotate with Pundit

Ts-213,406r[3]et407r[1]

Plain text - See in WittgensteinSource.org - Annotate with Pundit

Refers to:

Platon (3)
Frazer, Sir James (3)
Nietzsche, Friedrich: Jenseits von Gut und Böse (2)
Nietzsche, Friedrich (2)
Lichtenberg, Georg Christoph: Sudelbuch H 146 (2)
Lichtenberg, Georg Christoph (2)
Ernst, Paul: Nachwort zu Kinder- und Hausmärchen; III (2)
Ernst, Paul (2)
Tolstoj, Lew: Was ist Kunst (1)
Tolstoj, Lew: Tagebuecher 1885-1901 (1)
Tolstoj, Lew: Tagebuch; I (1)
Tolstoj, Lew (1)
Sraffa, Piero: Tba. (1)
Sraffa, Piero (1)
Spengler, Oswald: Der

Figure 20. Applying a second facet “Discusses : “ts-213 philosophy”.

He then selects this facet value, further restricting the search results, and discovers that it brings up several entities (persons, including philosophers) as possible values under the facet “refersTo” (figure 22).

By surveying this list of entities that the graph suggests the Bemerkungen “refers to”, the scholar uses his knowledge to come to the conclusion that Wittgenstein’s writings were influenced by “continental” rather than “analytical Anglo-Saxon” conceptions of philosophy. In addition, he assumes that studying / close reading of these sources will give him a clearer idea about the network of concepts Wittgenstein’s conception of philosophy is linked to. In other words, he looks at the keywords listed under the “refers to” facet and assumes that the “refers to” facet would lead him to concepts and philosophers that he would need to study further in order to get a better idea of Wittgenstein’s idea of philosophy. Moreover, by adding the facet value “Secondary Source” (which is equivalent to remove the first filter applied) the browser brings up new relevant entities such as articles that “discuss” Bemerkungen themselves containing a “discussion” of philosophy. He can now follow the Web resources linked to the philosophers, and immediately start learning about them as well. With this, the scholar has concluded his short experiment.

There are several elements of this small experiment, which can help us to understand how humanists would like to reason with Linked Data. These include the purpose of the WO itself, the research object and method of the experiment, and the conclusions made by the scholar.

As mentioned before, the creation of ontologies and vocabularies using Linked Data already entails a practice of reasoning, as scholars need them to contain the information and contingencies that will allow for further reasoning with the data. This means that having a look at how and for what purpose vocabularies and ontologies have been created can in and of itself be indicative of the kinds of reasoning scholars want to see enabled by Linked Data. This also means that ontologies and vocabularies can be seen as reflections of research questions and methods in the field and domain for which they have been created. With this in mind, the WO's general purpose, according to its creators is to provide other scholars with tools to assist them in their own research of Wittgenstein's Nachlass, including, among other things, a representation of the key concepts in the corpus and links to secondary sources that may help scholars understand the concepts and ideas expressed in the primary sources. The WO can be considered an attempt to create a knowledge representation or model of the research landscape concerned with Wittgenstein's Nachlass. The ontology appears to reflect on a research process of close reading and critical analysis of both primary and secondary sources. This process can be aided by the technology, but has to be accomplished by the researcher.

In contrast to the GEI vocabulary, the WO largely captures (explicitly states) only "factual" statements concerning the primary text and excludes, for example, the ontology-creator's interpretations of the content of the sources. Although choosing key concepts does involve close reading and a certain level of interpretation of the text, this process does not attempt to definitively or explicitly state the meaning of the concept in the text for Wittgenstein research, but merely point to the fact that certain constellations of keywords and sources relevant for scholars are considered to be linkable with certain texts. The creators assume that each scholar will want to partake in close reading and meaning-making using all available sources as well. This indicates that, in the humanities, the meaning to be made from text is variable and dependent on not only the content, but also on the interaction of the individual researcher with the content and with other researchers as well. The WO provides the researcher with a tool to aid this interaction.

The purpose of the ontology is to support the research question explored by the scholar, which was to come to an understanding of Wittgenstein's conception of philosophy by looking at the key concepts in the text (as catalogued in the WO). The scholar bases this approach among other things on the premise that an overview of the key concepts can help him to gain a quick orientation in the text and to therefore understand the information better. This research question is therefore not only focussed on content, but on a meta-level on the research process as well: He would like to discover if an overview of keywords can tell him something about the content.

The accompanying research process described by the scholar does lead to an answer on both levels. As in the GEI experiment, the scholar is able to obtain new information based on comparing the results list with his real-world knowledge. Although this information is not explicitly contained in the triples, the scholar notices that all of the philosophers "referred to" in the "Bemerkung" in which Wittgenstein discusses his concept of philosophy have something in common: they were so called "continental" philosophers. This is new information for the researcher. He has therefore learned something about Wittgenstein's concept of philosophy. At the same time, the scholar is called to carry out his own close reading, as he might not yet be sure of what these philosophers have said. On the other hand, as the label of these philosophers as "continental" was not contained in the WO and is new information coming *from* the researcher, this information is a valuable addition to the WO and can be recorded in the form of triples via Pundit.

Scholar Two: Wittgenstein's critique of picture theory

The second experiment on the ontology of the WAB was carried out by another scholar who can be considered to be an expert in both Wittgenstein research and ontologies in general. Although not intimately familiar with all of the details of the WO, he is currently working, among other things, on an ontology to represent certain concepts in Wittgenstein's *Tractatus*. On one level, this experiment proved to be unsuccessful, as the scholar was unable to use the WO to come to any new conclusions about the data. In fact, he was disappointed with his experience using the ontology browser. This however, led to fruitful discussions with the ontology's creator and the DM2E working group, which will be discussed below. The experiment will therefore be briefly described in this section.

The scholar's initial research question was to use the faceted browser to find what Wittgenstein wrote on the picture theory in the Big Typescript (TS-213). His first step was to enter the following question in the search bar of the ontology browser in natural language. He entered: "What is Wittgenstein's critique of the picture theory in Ts-213?" He then realised that the search bar does not work using natural language.

His second step was then to only enter the expression "picture theory"; he retrieved five results from the secondary sources. In a fourth step he then entered only the expression "picture" and the browser suggested, in his words "many completions or additions". He then picked the facet "Ts-213-021 Similarity of sentence and picture" and got 12 results. Being a Wittgenstein scholar, he knew that these results were correct and that he could follow them to Wittgenstein Source and read the German text.

His own conclusion was that he could not do very much with the ontology: "This is all I can do with the ontology?"

Results

We invited the second Wittgenstein scholar to discuss his difficulties using the WO explorer with the creator of the WO and documented the resulting discussion. The second scholar was able to uncover certain weaknesses in the implementation of WO on the facets browser as also the underlying dataset. The discussion with the scholars also revealed some basic issues involved in the reasoning process with faceted browsers.

One major issue that was raised in the discussion revolves around trust and authority. As was seen in the GEI experiment, in order to be able to make inferences about the data, humanities scholars need context. This not only includes the context of the dataset itself, but of its creation as well. In this regard, the second Wittgenstein scholar remarked that he would have needed an explanation of the ontology included in the browser in order to be able to understand it fully. The full extent of the underlying dataset was not immediately evident, and the scholar was irritated by the fact that Wittgentein's *Tractatus* was not included.[124] Seeing as the modelling practice can be related to the scientific method, this includes knowing who made it, for what reason and using what methods and principles. The basis of any further analysis by the scholar relies on the data contained in the ontology being accurate, adequate and, to a certain extent, authoritative.

For this reason, the scholar suggested using a persistent identifier such as the one established in ORCID[125] for each entry in the ontology to identify its creator. Of course, Pundit solves this problem to some extent using the notebook system, but this shows that scholars would like to be aware of this at every step of the process.

One reason mentioned for needing such a persistent identifier of the ontology creator is the division of knowledge in science. It was argued that even Wittgenstein scholars are often experts in only one area such as "religion and Wittgenstein". Knowing which scholar is responsible for which information increases transparency; and the knowledge explicated by one scholar for a particular topic will have more weight or authority than another. For an ontology to have stability and authority, it would ideally need area editors.

[124] Due to legal issues.
[125] http://orcid.org/

Another topic discussed in the expert interview revolved around the process of ontology creation itself, which was just as important of a research object for our scholars as Wittgenstein scholasticism. They discussed the need for documentation and standardisation of not only the ontology design, but also of the creation process, so that other scholars or perhaps even machines can understand and recreate it. Both scholars believed that each scholar should be able to question and explore the design of the ontology itself. In this context they saw the ontology browser as a medium of communicating one particular view of the domain, which could form the basis of ontological comparison, implicitly as well as explicitly.

6.3 Conclusion

This chapter primarily set out to inquire “the kinds of reasoning humanities scholars want to see enabled with Linked Data” (DoW). For this purpose, we wanted to foster a perspective on reasoning that is not focused on the aspects computer science is predominantly concerned with, and take into account the prerequisites for the use of Linked Data in the context of interpretative research in the humanities. We therefore proposed the concept of “interactive reasoning” as an attempt to approach “reasoning” as a scholarly practice in the context of the Semantic Web.[126] In contrast to automatic inference by machines in the Semantic Web, the term “interactive reasoning” stresses the intention to facilitate reasoning practices for humanists, who conduct their research in the context of Linked Data applications. In this context we specifically concentrated on how humanists can use faceted browsers to explore and reason with Linked Data.

Our method for explaining this issue involved working with three humanities scholars on two particular use cases. In the first use case, we asked a historian to create a small vocabulary for the purpose of semantically annotating a specific corpus of historical textbooks. He then used the Pundit Search faceted browser to query the graph he had created looking for answers to a particular research question. In the second use case, we applied a faceted browser to an existing ontology that was created to be a representation of the research landscape of Wittgenstein’s Nachlass published on Wittgenstein Source. We then had two different Wittgenstein scholars attempt to answer research questions about the Nachlass using the faceted browser. Our first scholar was intimately familiar with the WO, our second scholar was an expert in both Wittgenstein and in the topic of ontologies, but was much less familiar with the WO itself.

It was important for us to explore how humanities scholars understand and explore the graph, in particular applications like the Wittgenstein Ontology Explorer to visualise parts of the underlying structured data. And therefore provide a means to the scholar to engage with that structure and subsequently to contribute to it. What Hitchcock (2013) said about the effects of Googling in the field of history also applies to our topic of reasoning and Linked Data: You need to understand what is going on with the graph and how you obtained the results. When scholars create and apply their own data it gives them the necessary context to understand the result.

The experiments[127] were conducted in order to complement the theoretical research on the functional requirements for the translation of Scholarly Operations[128] as well as on the possible application of “reasoning” for scholarship in the humanities in the context of Linked Data with an empirical perspective. The observations uncovered a common threefold structure of the translation and application of interpretative scholarly practices and

[126] Cf. the Scholarly Domain Model, esp. Interpretative Modelling.
[127] Also cf. “Report on Experiments”.
[128] Cf. the Scholarly “Domain Model.

reasoning to a Linked Data application environment. The first phase includes the initial creation or reuse of vocabulary, to represent the knowledge about certain aspects of the research domain and the methodological approach, to guide and structure the subsequent annotation process. If you understand this first modelling process, you also understand how the researcher plans to reason with the data. The second phase involves the process of interaction with the corpus data, deciding which resources and entities should be annotated and which semantic annotation apply. The third phase involved querying the resulting graph and making inferences about the data that had been created. Therefore, we suggest the following principal and formal three phases of reasoning in Linked Data context:

1. Conceptualising: Vocabulary selection, modification, or creation which includes the translation of a research interest or question into a Linked Data conceptualisation such as an annotation vocabulary. The vocabulary formalises and explicates the "reasoning" result of a first genuine part of the research process which is based on assumptions or hypotheses about the research to be conducted.

2. Annotation: The application of the annotation vocabulary to research objects by creating annotations, in the current context with Pundit. Working with the actual texts probably is most commonly associated with the actual research conducted and involves close reading and interpretation, here expressed and formalised as annotation triples.

3. Exploration: The assessment of the created triples by visualisation, in the current context in a faceted browser which we consider a "low-hanging fruit" for applying one's own reasoning practice to a given knowledge base. Here the researcher explores the previous reasoning and creates new hypotheses which may feed back into the annotation vocabulary and may initiate a new annotation phase.

When analysing our use cases, we noticed that the ontology browser allowed the scholars to quickly gain an insight into certain groups of Linked Data, made by picking certain facets that combined triples of data from a number of sources. The researchers were then able to make inferences about these results, comparing their observations with textual, intertextual and real-world knowledge: the scholar was able to observe, for example, patterns and anomalies implicit in the data. These observations led to the potential creation of new information and therefore new triples. In the GEI use case, our scholar discovered that the historical German textbooks chose to portray the German Empire as a peaceful country. This was done by not only emphasising the topos of peace, but also by using "national" topics to characterise German. In contrast, other countries were painted as being aggressors. The concept of "national" was new information that could have been added to the corpus to make this new knowledge explicit. In the Wittgenstein Ontology use case, our scholar came to the conclusion that Wittgenstein's philosophy is influenced by "continental" as opposed to "anglo-saxon" philosophers. This newly inferred information could potentially also be added to the Linked Data vocabulary.

In addition to uncovering one method of coming to conclusions about data using faceted browsers, the experiments also uncovered two very important aspects of reasoning in the digital humanist realm. The GEI use case showed us that context is extremely important. When modelling the domain of German historical school books, GEI attempted to incorporate as many historical facts as possible about the works. This context included not only dates and places, but also facts that might help to uncover biases and values, such as the religious affiliation of the authors. This played a large role in his vocabulary, which tried to make certain value concepts explicit. In the WO use case, the conversation between our scholars revealed that the reasoning process relies heavily on the context of the research itself. This means scholars need to know who is creating what annotations for what purposes and how. As a result of our use cases, we have determined that context on all levels needs to be taken into consideration when considering reasoning scenarios with Linked Data in the digital humanities.

In conclusion, the three step process discussed here represents the basic structure of one way of how humanists do and want to reason in the context of Linked Data and interpretative approaches. The same basic iterative 3-tiered process has been identified in other experiments we conducted. Of course, Pundit predetermines the outcomes of the experiments to a certain extent in terms of the available functionality. However, we nevertheless found this three step procedure as fruitful to represent simple interpretative approaches from the humanities in a Linked Data context. All experiments demonstrated its potential. In that regard, we propose further systematic research be conducted based on this principle approach in order to deepen our understanding of how humanists do and want to reason, and more generally conduct parts of their research, in the Linked Data and Semantic Web context.

Acknowledgments

We would like to thank the following persons for their support in the experiments and their contributions to this chapter: Alessio Piccoli (Net7), Andreas Weiß (GEI), Jakub Mácha (Masaryk University, Czech Republic)

6.4 References

- **Beynon, M.; Russ, S. and McCarty, W.** (2006). Human Computing: Modelling with Meaning, Literary and Linguistic Computing, 21 (2), pp. 141–157: 10.1093/llc/fql015 (accessed 31 January 2015).
- **Bowen, W. R. and Siemens, R. G.** (eds) (2008). New technologies and Renaissance studies: Contributions from annual conferences held 2000 to 2005 during gatherings of the Renaissance Society of America. Tempe, AZ: Iter.
- **Blanke, T. and Hedges, M.** (2013). Scholarly Primitives Building institutional infrastructure for humanities e-Science, Future Generation Computer Systems 29 (2): 654–661: 10.1016/j.future.2011.06.006 (accessed 31 January 2015).
- **Crombie, A. C.** (1994). Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts. London: Duckworth.
- **Deegan, M. and Sutherland, K.** (2009). Transferred illusions: Digital technology and the forms of print. Farnham: Ashgate.
- **Dentler, K. et al.** (2011). Comparison of Reasoners for large Ontologies in the OWL 2 EL Profile. In Semantic Web 1 (1-5). http://www.semantic-web-journal.net/sites/default/files/swj120_2.pdf (accessed 31 January 2015).
- **Eiter, T. and Krennwallner, T.** (eds) (2012). Reasoning Web. Semantic Technologies for Advanced Query Answering: Proceedings of the 8th International Summer School, Vienna, Austria, September 2012. Berlin: Springer. (Lecture Notes in Computer Science, 7487).
- **Gradmann, S.** (2013). Semantic Web und Linked Open Data. In Kuhlen, R., Semar, W. and Strauch, D. (eds), Grundlagen der praktischen Information und Dokumentation. Handbuch zur Einführung in die Informationswissenschaft und -praxis. Berlin: De Gruyter Saur, pp. 219–228.

- **Grassi, M. et al.** (2012). Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries. In Mitschick, A. et al. (eds), Semantic Digital Archives 2012: Proceedings of the Second International Workshop on Semantic Digital Archives, Paphos, CY, September 2012. http://ceur-ws.org/Vol-912/paper4.pdf (accessed 31 January 2015).

- **Grassi, M. et al.** (2013). Pundit: augmenting web contents with semantics. Literary and Linguistic Computing 28 (4): 640–659: DOI: 10.1093/llc/fqt060 (accessed 31 January 2015).

- **Gruber, T. R.** (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing, International Journal of Human-Computer Studies 43 (4-5), pp. 907–928.

- **Hacking, I.** (1985). Styles of Scientific Reasoning. In Rajchman, J. and West, C. (eds), Post-analytic philosophy. New York: Columbia University Press, pp. 145–164.

- **Hitchcock, T.** (2013). Confronting the Digital, Or How Academic History Writing London the Plot. In Cultural and Social History 10 (1): 9-23: 10.2752/147800413X13515292098070 (accessed 31 January 2015).

- **Holyoak, K. J. and Morrison, R. G.** (eds) (2012). The Oxford Handbook of Thinking and Reasoning. Oxford: Oxford University Press.

- **Kuhlen, R.; Semar, W. and Strauch, D.** (eds) (2013). Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und -praxis. Berlin: De Gruyter Saur.

- **Ludwig, S. A.** (2010). Comparison of a Deductive Database with a Semantic Web Reasoning Engine. Knowledge-Based Systems 23 (6): 634–642: 10.1016/j.knosys.2010.04.005 (accessed 31 January 2015).

- **McCarty, W.** (2005). Humanities computing. Houndmills: Palgrave Macmillan.

- **McCarty, W.** (2012). The residue of uniqueness, Historical Social Research 37 (3): 24–45. http://www.jstor.org/stable/i40077811 (accessed 31 January 2015).

- **McCarty, W.** (2008). Being Reborn: The Humanities, Computing and Styles of Scientific Reasoning. In Bowen, W. R. and Siemens, R. G. (eds), New technologies and Renaissance studies. Tempe, AZ: Iter. (New technologies in medieval and renaissance studies, 1), pp. 1-22.

- **Meister, J. C.** (ed) (2012). Digital Humanities 2012: Conference Abstracts, University of Hamburg, July 16-22. Hamburg: University Press.

- **Oldman, D., Doerr, M. and Gradmann, S.** (n.d.). ZEN and the Art of Linked Data. New Strategies for a Semantic Web of Humanist Knowledge. To be published in Schreibman, S., Siemens, R. G and Unsworth, J. (eds), A new Companion to Digital Humanities. Oxford: Blackwell [preprint].

- **Pesce, M.** (1999). SCOPE1: Information vs. Meaning. http://hyperreal.org/~mpesce/SCOPE1.html (accessed 31 January 2015).

- **Pichler, A. and Zöllner-Weber, A.** (2012). Sharing and debating Wittgenstein by using an ontology [doi: 10.1093/llc/fqt049]. In: Literary and Linguistic Computing, Vol. 28 (2013) / Number 4. pp. 700-707. (UK) Oxford: Oxford University Press.

- **Rajchman, J. and West, C.** (eds) (1985): Post-analytic philosophy. New York, NY: Columbia University Press.

- **Svensson, P.** (2009): Humanities Computing as Digital Humanities, In Digital Humanities Quarterly 3 (3). http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html (accessed 31 January 2015).

- **Zoglauer, T.** (2008): Einführung in die formale Logik für Philosophen. Göttingen: Vandenhoeck & Ruprecht.

- **Zöllner-Weber, A.** (2009): Ontologies and Logic Reasoning as Tools in Humanities?, In Digital Humanities Quarterly 3 (4). http://www.digitalhumanities.org/dhq/vol/3/4/000068/000068.html (accessed 31 January 2015).

6.5 Appendix: Guidelines

Guidelines for the documentation of DH reasoning practices in the context of DM2E and the tools Pundit and the faceted browser in Ask

The principle goal of this experiment is to observe the information behaviour of humanists while working with Linked Data. The technical and conceptual scope of the experiment is a faceted browser which allows to explore triple data. The particular research interests focuses on the work and reasoning process the humanist applies within this particular setting while trying to find answers to particular research questions relevant to their domain of discourse.

The outcomes of this experiment will provide empirical evidence for the type of reasoning humanists want to apply to triple data.

The participants are asked to choose ***two*** *research questions which are* ***relevant*** *for their particular domain of discourse and which they expect to be* ***applicable*** *to the corpus at hand!*

The participants will then try to answer these research questions using the faceted browser and create a self-documentation of the technical and reasoning procedure they applied.

The documentation should address the following four sections: The first section provides a brief overview on the characteristics of the corpus (i.e. the data) you are using for the experiment. The second section gives a brief description for each research question you will apply to the corpus. The third section provides a guideline for recording your work process for each research question in the faceted browser. The fourth section summarizes your experiences and considers proper reasoning scenarios by answer the following questions:

1. Corpus

Please provide a short overall description of the corpus and its Linked Data representation you are using!

- Who created the corpus, for whom and for what purpose?

- What kinds of data does it contain (e.g. annotations, ontology, textual data, digitized images etc.)?

- Who created the linked data representation of the corpus, for whom and for what purpose?

- Describe the faceted browser! What kinds of entities and relations did you mark-up in Linked Data (what vocabularies and ontologies did you use) to create your facets?
- How many triples were incorporated in the faceted browser

2. Research questions

For each research question please provide a brief general characterisation!

- Why did you choose this particular research question? What is the relevance of the research question for the particular domain of discourse your corpus is addressing?
- Which answers do you expect and why do you think the corpus will provide sufficient information?

3. Protocol

For each research question create a step-by-step protocol ("Verlaufsprotokoll") of each step you are taking during your work with the faceted browser. This protocol should also support you in writing up a summary of your process and to assess and justify the steps you took. The main point is that you try to be self-conscious about which kinds of assumptions and conclusions you are drawing along the way.

For each step, try to describe in a detailed manner each single action/step you took and then try to explain why you performed the action/step (consider the guiding questions!).

Describe the process!	**Reflect on the Process!**
Research question / problem:	
Step/Action 1:	
Step/Action 2:	
Step/Action 3:	
...	

4. Summary

(a) For each research question please provide a summary of your process by considering the following questions.

- What is the final answer you found for the research question?
- Assess the quality and usefulness of the answer you found from a scientific viewpoint!
- Compare (a) the reasoning process, or aspects of this process, you applied during your work with the faceted browser to (b) the reasoning process you would apply when working in a mostly non-digital setting.

(b) For each research question, please provide a reflection on the reasoning process involved.

- Did the premises you had before starting working on the research question influence the way you reasoned or proceeded? If so how?
- How would you describe the research method/process used in answering the research question?
- How did the assumptions you made about the data and the research methods inform the conclusions you came to?
- Do you presume the results to be trustworthy or do you have doubt about their trustworthiness?

(c) As an overall consideration please provide your viewpoint on the potential of proper reasoning and inference scenarios for your particular use case.

- Discuss the potential of reasoning software for answering the particular research question!
- Did any new questions arise from the results generated?
- Which conceptual or technical aspects you encountered while using the faceted browser influenced your work and how?

7 Conclusion and Recommendations

The results of the Task 3.4 stem from the working group's extensive theoretical and empirical research to explore the functional primitives in the digital humanities using DM2E content and tools. In the following, we summarise some of the main findings including recommendations for future work on Digital Humanities Virtual Research Environments (VRE) in Linked Data contexts, such as Europeana.

The Scholarly Domain Model can be seen as an example of documenting the recursive/iterative modelling process necessary to capture and articulate how scholars conduct research in the real world. We therefore would argue that, when developing functionalities and tools for Digital Humanities projects, scholars and computer scientists should consult a model such as the SDM in order to increase the sustainability of the Virtual Research Environment. The Scholarly Domain Model is an explicit but not definite model that is open and extensible, thereby easily adaptable to different use cases. The "functional primitives" resemble the four levels of abstractions for the constituents of the scholarly domain, identified during our work. These are the Areas, Scholarly Primitives, Scholarly Activities, and Scholarly Operations for modelling the scholarly domain on the basis of the practices of digital scholarship in the humanities. In particular, the "types of operations" have been conceptualised as the Scholarly Operations stressing the importance of continuous translation and modelling with strong connection to the scholarly practices. In this sense, the SDM constitutes a framework for sustainably modelling of digital scholarly practices. The SDM RDFS/OWL representation is a starting point for the implementation of the SDM in a monitoring context.

The main outcome of the experiments and the reasoning task is the recognition of a basic tripartite research and reasoning process revolving around the Scholarly Activities of Interpretative Modelling and Annotating in the context of Linked Data based VREs. It involves first creating or choosing an annotation vocabulary, applying the vocabulary to the source material, and exploring the created annotations in order to create new hypotheses. This tripartite process can be conceptualised as an expression of research processes on the level of the Scholarly Operations and described using terminology from the SDM. Encouraging humanists to work with Linked Data requires taking a step towards translating the objects of study and methods of humanist research in Linked Data paradigm and taking a step away from concentrating on what can be automated: The tripartite process can be understood as the principle answer to the question which kinds of reasoning humanists want to see enable in Linked Data based VREs: The process enables them to apply their own reasoning practices to a certain extent and also provides a framework for further systematic investigation into the question how interpretative approaches of humanists might translate and be applied in Linked Data annotation environments.

The reasoning and interpretive modelling processes of each stage of this tripartite process can be described using terminology from the SDM. For purposes of illustration, we have chosen Conceptualising, Annotating and Exploration. First, conceptualising, referring here to the mental process underlying vocabulary selection, modification, and creation, includes the translation of a research interest or question into a Linked Data conceptualisation such as an annotation vocabulary or ontology. The vocabulary formalises and explicates the "reasoning" result of a first genuine part of the research process which is based on assumptions or hypotheses about the research to be conducted. The second step involves annotating, i.e. the application of the annotation vocabulary to research objects by creating annotations, in the current context with Pundit. Working with the actual texts involves close reading and interpretation, here expressed and formalised as annotation triples. The third stage allows for exploration, which includes the assessment of the created triples by visualisation, in the current context in a faceted browser which we consider a "low-hanging fruit" for applying one's own reasoning practice to a given knowledge base. Here the researcher explores the previous reasoning and creates new hypotheses which may feed back into the annotation vocabulary and may initiate a new annotation phase.

Furthermore, our research can support the current opinion in the literature which states that the research in the humanities is often concerned with meaning, and therefore cannot necessarily be adequately represented by fully automated reasoning processes. Therefore, useful application of reasoning in Linked Data based Virtual Research Environments (VREs) for digital scholarship in the humanities requires further investigation and consideration of the representation of values and belief systems, as well as the structure of argumentation, using RDF(S) vocabularies. For example, vocabularies that can express opinions, values and beliefs as well as provide reasons for statements have been one of the most demanded features during the experiments. This constitutes a prerequisite for the application of humanistic methods as well as the appropriate representation of the research objects in the process of humanities scholarship.

8 Appendix: Related Publications and Presentations

Publications

- Stefan Gradmann, Julia Iwanowa, Evelyn Dröge, Steffen Hennicke, Violeta Trkulja,Marlies Olensky, Christian Stein, Alexander Struck, Konstantin Baierer (2013) Modellierung und Ontologien im Wissensmanagement. Information Wissenschaft & Praxis 2013. http://dm2e.eu/files/iwp-2013-0016.pdf (accessed 31 January 2015).
- Schreibman, S., Gradmann S., Hennicke S., Blanke, T., Chambers, S., et. al. (2013). Beyond Infrastructure – Modelling Scholarly Research and Collaboration. In: Digital Humanities 2013: Conference Abstracts, University of Nebraska–Lincoln, USA, 16-19 July 2013. pp. 386-389. http://dh2013.unl.edu/abstracts/files/downloads/DH2013_conference_abstracts_print.pdf (accessed 31 January 2015).

Presentations

- Medieval Cultures on the Web, Italy, 07/03/2012, Stefan Gradmann, Christian Morbidoni.
- Data Modelling the Humanities, Brown University, US, 14/03/2012, Stefan Gradmann (http://bit.ly/1bRiaN7).
- Digital Humanities, Luxembourg, 20/03/2012, Stefan Gradmann.
- Digital Humanities Panel, Semantics and the Web, France, 20/04/2012, Stefan Gradmann (http://bit.ly/GNCJf6).
- Conference of the Libor Manuscripts Section, France, 30/05/2012, Stefan Gradmann.
- Leipzig eHumanities Seminar, Germany, 07/11/2012, Stefan Gradmann.
- Easy Tools for Difficult Texts, Netherlands, 19/04/2013, Stefan Gradmann.
- CERN Workshop on Innovations in Scholarly Communication (OAI8), Switzerland, 19/06/2013, Stefan Gradmann, Alessio Piccioli.
- APEX-Conference, Republic of Ireland, 26/06/2013, Steffen Hennicke (http://slidesha.re/19zN6ft).
- Digital Humanities Conference (DH2013), US, 19/07/2013, Stefan Gradmann, Susan Schreibman.
- DH Case: Collaborative Annotations in shared environments (Demo tools, Data models and digital editions), Italy, 10/09/2013, Christian Morbidoni, Simone Fonda, Alois Pichler.
- EVA/Minerva Jerusalem International Conference on Digitisation of the Culture, Israel, 10/11/2014, Kristin Dill (http://slidesha.re/1rT4k5B).
- Semantic technologies for research in the humanities and social sciences (STRiX), Sweden, 24/11/2014, Kristin Dill, Steffen Hennicke, Gerold Tschumpel, Christian Morbidoni, Alois Pichler and Klaus Thoden.