Classical philology, text-based and Holocaust studies facing digital research infrastructures: from practice to requirements

Stavros Angelis¹, Agiatis Benardou¹,³, Panos Constantopoulos¹,², Costis Dallas¹,³,⁴

¹ Digital Curation Unit – IMIS, Athena Research Centre
G. Bakou 17, GR-11524 Athens
Greece

² Department of Informatics, Athens University of Economics and Business
76 Patission str., GR-10434 Athens
Greece

³ Department of Communication, Media and Culture, Panteion University
136 Syngrou Av., GR-17671 Athens
Greece

⁴ Faculty of Information, University of Toronto
140 St. George Street, M5S 3G6 Toronto, Ontario
Canada

s.angelis@dcu.gr, a.benardou@dcu.gr, p.constantopoulos@dcu.gr, c.dallas@dcu.gr

Abstract

This paper reports on research on scholarly research practices and requirements conducted in the context of the Preparing DARIAH and, subsequently, the EHRI (European Holocaust Research Infrastructure) EU-funded e-infrastructure projects. It summarises the findings of earlier research, and describes the plan of work and methodology of the empirical research designed and conducted by DCU in the context of both projects. In DARIAH, this was based on open-questionnaire interviews with arts and humanities researchers, and later formalised into a conceptual model for scholarly research activity, before presenting the finalized user requirements recommendations. In EHRI, DCU is following a mixed methods approach: the first, qualitative stage is currently under way, consisting of semi-structured interviews with a representative selection of Holocaust scholars; the second, confirmatory stage of this research plan consists of an online questionnaire survey, composed mostly of closed questions amenable to statistical summarisation and hypothesis testing. Finally, this paper presents plans for future work on user requirements for EHRI.

Introduction

The growing importance of digital infrastructures for arts and humanities research, witnessed in the emergence of several international and discipline-based initiatives in Europe, North America and Oceania, highlights the need to establish reliable knowledge on scholarly requirements such infrastructures purport to satisfy. This paper presents the results and current progress of research conducted during the last three years by the Digital Curation Unit-IMIS, Athena Research Centre, Athens, in the field of scholarly information work and user requirements for digital infrastructures, and in the context of two EU-funded projects: Preparing DARIAH, the preparatory phase of the European digital infrastructure in the arts and humanities (2008-2011) and the European Holocaust Research Infrastructure (EHRI) initiative (2010-14).

Requirements work in the two projects has been different as regards the scope of disciplinary perspectives they cover, as well as in workplan and particular methodologies used. Preparing DARIAH covered a broad range of humanities disciplines, including archaeology, Classics, history, history of art, philosophy, ethnomusicology and social/cultural anthropology, in line with the broader mandate of the project to define strategies, operational models and functional specifications addressing the needs of all these epistemic communities; the project was designed to produce specifications, which are now taken forward to the implementation stage, planned to start in 2012 under the responsibility of a European Research Infrastructure Consortium (ERIC) formed for this purpose. EHRI, on the other hand, focuses on the highly interdisciplinary field of Holocaust studies, covering epistemic perspectives as diverse as history, archival research, and visual studies, and involves the actual development and operation of a thematic research portal integrating research repositories and search functionalities, and a Virtual Research Environment. In addition, EHRI builds on the methodological framework and substantive results of Preparing DARIAH, with the explicit goal of confirming their validity in the field of Holocaust research, and of identifying specific patterns of scholarly work and user requirements that are particular to this field. In both projects, an important goal has been, among others, to advance knowledge on aspects of information use and infrastructure requirements related to philological and text-based scholarly research and sources.

Background

Systematic studies of scholarly information activity date from the 1990s. In an important early study, Bates and her team investigated information work of humanities scholars visiting the Getty Research Institute. Results include establishing the major importance of named entities (proper names, places, titles of works) as access points for resource discovery, but also the fact that the introduction of digital tools did not supplant traditional methods but co-existed in tandem with them (Bates, Wilde, & Siegfried, 1995); the study focused on the needs of humanists with regard to scholarly objects, such as those found in a library or an archive. Other researchers investigated information aspects of scholarly use of primary and secon-
dary sources, mostly focusing on text-based disciplines, such as history (Dalton & Charnigo, 2004; Delgadillo & Lynch, 1999; Duff, Craig, & Cherry, 2004; Tibbo, 2003). More recently, short reports, based on the consultations and research conducted by the AHRC Methods Network in the UK, summarise needs and likely scenarios for ICT use in humanities research in particular fields, including history, art history, archaeology, and museums and heritage; also, a conceptual overview of ICT research methods employed by researchers has been developed by the Methods Network in the form of a thesaurus (“methods taxonomy”).

Conceptualisation of different kinds of scholarly information activity has been a major interest of researchers working in the field of user information behaviour. Ellis and his team at Sheffield University had used a grounded theory approach to study typical information processes of four different disciplinary communities across the sciences, social sciences, and the humanities (Ellis, 1993). They identified six common processes across disciplines: starting, chaining, browsing, differentiating, monitoring, and extracting. These were supplemented a decade later by Meth and Tibbo (2003) with further processes of accessing, networking and verifying. Brockman et al. presented a related broadly based conceptual framework of the nature of scholarly work, focusing on the text-oriented processes of reading, collaborative networking, researching and searching, and ways of writing, and emphasizing the particularities of information work in the humanities vis-à-vis other disciplines (Brockman, Neumann, Palmer, & Tildine, 2002).

The notion of “scholarly primitives”, initially advocated by Unsworth (2000) in the then fledgling field of humanities computing with reference to processes employed by literary scholars, was later found to be useful to identify common, low-level scholarly activities in other humanities disciplines, both text-based and not, such as history and historical geography (Mostern, 2006). A recent synthesis by Palmer et al. identifies five broadly conceptualised “scholarly activities”: searching, collecting, reading, writing, and collaborating. These, as well as a bucket of “cross-cutting primitives” are further refined to a more detailed, and useful, list of twenty granular “scholarly primitives” (Palmer, 2010; Palmer, Teffeau, & Pirmann, 2009). Palmer et al. reached important conclusions on the relative importance of specific scholarly information activities: browsing, collecting, re-reading, assembling, consulting and notetaking were found to be particularly common in the humanities, while chaining, accessing, assessing, disseminating and networking were seen as equally applicable to the humanities as well as other disciplines. Chaining, in particular, was identified as the most notable activity among humanists as they seek information. In addition, probing and translating activities were found to be most common in interdisciplinary research, a noteworthy finding, considering the frequently interdisciplinary nature of work in the humanities.

Scholarly information behaviour research is relevant to the context of designing appropriate digital information infrastructures for scholarship. A statement of strategic direction with regard to the digital infrastructure for the humanities in the US, the 2006 report of the American Council of Learned Societies Commission on Cyber infrastructure for the Humanities and Social Sciences, provided a useful consensus view on what constitutes scholarly practices that need to be addressed by ICT infrastructure and tools (ACLS, 2006). The lineage of scholarly primitives was taken forward through the identification of four groups of scholarly information activities (discover, gather, create and share) by a study aimed at defining appropriate infrastructures and services at the University of Minnesota Libraries (2006). A useful broader framework for conceptualising the relationship between disciplinary practices in the humanities, and the use of documents and data for scholarly research, with technological infrastructures and tools is provided by Borgman (2007).

**Establishing scholarly information requirements in Preparing DARIAH**

Understanding the nature and information requirements of scholarly research in the Arts and Humanities emerges as an important motivation and a prerequisite for the definition of affordances of digital infrastructures, services and tools fit for the purpose of current and future scholarly research, notwithstanding existing differences between disciplines, research fields and methodological approaches. Requirements need to be substantiated by evidence drawn from scholarly practice, and encompass not only which digital resources, services and tools humanities researchers use, and how, but, more generally, how they interact with the whole spectrum of information and conceptual entities – digital as well as non-digital – involved in the research process. Apart from primary research data and bibliographic sources, these include finding aids, ways of organizing and sequencing research activities, terminological and encyclopaedic resources, standard procedures, and tools.

Towards this goal, as part of the Preparing DARIAH EU project, the DCU designed and conducted in 2009 a programme of twenty-four interviews with arts and humanities scholars from across Europe (Benardou, Constantinopoulos, Dallas, Gavrili, 2010). Interviewees were scholars at various levels of career advancement, members of academic departments or research institutes, engaged in research across a spectrum of different disciplines in the arts and humanities.

All interviews were recorded with the consent of the participants, transcribed into written form, segmented and tagged on the basis of conceptual content. The duration of the interviews ranged depending on the interviewees’ personal interests, their research methodology, and their willingness to contribute with reports of other activities which we considered would be relevant to our inquiry. Interviews were based on a semi-structured thematic repertoire but were open and conversational in style, encouraging the elicitation of further information when this was justified by the scope of the present research. Despite known difficulties arising in the analysis of free text interviews, the semi-structured format and open style adopted helped identify important differences between specific methodological perspectives and thematic interests within and across disciplines.

In order to help the elicitation of useful discussion on the findings of our research, we developed several different scenarios of use, representing typical configurations of scholarly activities involving the interaction between scholars and resources useful for their research. For communication purposes, one of these user scenarios has been represented in the form of a comic strip (Figure 1).
The findings of our analysis of scholar interviews reinforce the importance and increasing need that humanities scholars attribute to organizing information resources they collect and/or use during their research. They confirm that primary research materials in the humanities are collected amongst a wide range of heterogeneous sources, in the form of texts, images, or artefacts. Collections may be quite diverse, as they rely heavily on data recorded in personal notebooks filled with annotations, facts and references, and collected cultural artefacts, such as items that relate to the time, place, or object of study. Resources useful for scholarly research are often configured as structures of heterogeneous, complex, and densely interconnected objects, their intellectual content spanning the boundaries between categorical and factual knowledge, the ontological domains of the empirical, the actual and the real, and the epistemological divide between primary data and scholarly objects.

Personal collecting and archiving of research information—a still poorly understood phenomenon—ranging from primary data to a variety of scholarly objects, emerges as a widespread reality among humanities researchers. This practice precedes technology inception, but is certainly reinforced and further developed in the contemporary context of digital resources, tools and services. Indeed, personal collections and databases emerge among the most important sources of scholarly information for arts and humanities researchers. To be useful, primary and secondary research resources are discovered, gathered, appraised, digested, organised, categorised and stored for present and future scholarly use. Information organisation schemes are often the outcome of personal effort, while kinds of arrangement and storage of physical sources may vary from piles on the floor to structured file folder systems, and personally designed and implemented databases, varying in sophistication according to their creator’s familiarity with ICT tools. The specification of the structure and affordances of bespoke databases, and the procedures established for information management, rely heavily on established information seeking, annotation, versioning, and knowledge making practices.

Our analysis revealed that curation activities, i.e., those concerning the preservation, intellectual organisation, knowledge enrichment and evolution of information objects (Constantopoulos & Dallas, 2008), are of paramount importance to scholarly work, on a par with information-seeking activities typically given prominence in information behaviour research. In particular, annotation emerged as an important activity, regularly used by scholars to relate a concept, syllogism or hypothesis to a particular area of an image or a text passage, in a number of different ways. Given the evidence and user conceptualisation of annotation practice manifested in our interviews, user requirements could well be served by simple tools amenable to informal use, for instance supporting the superimposition of arrows and circles over a visual resource, or supporting the collaborative accumulation of layers of links connecting different kinds of information entities, and thus enabling alternative kinds of scrutiny and research approaches.

In attempting to draw useful conclusions from our interview research, we adopted a requirements-push rather than a technology-pull approach: i.e., we took the view that digital research infrastructures should seek to address the needs emerging from scholarly practice in the arts and humanities, rather than dictate to scholars how they should conduct their research. On this basis, our analysis led to the specification of ten user requirement recommendations (Table 1), published as part of the DARIAH official technical report.

| USER REQUIREMENT RECOMMENDATIONS (URRs) |  |
|----------------------------------------|  |
| URR1 | Extend the scope of infrastructure to encompass data and resource repositories on the one hand, and digital libraries of scholarly objects on the other. |  |
| URR2 | Develop appropriate identity and reference mechanisms. |  |
| URR3 | Provide support for aggregated and complex information objects. |  |
| URR4 | Provide effective representations, at the level of metadata and also access/management processes, for IPR protection and access amenable to a diversity of approaches and regimes (from open content to fully controlled IPR). |  |
| URR5 | Provide a semantic interoperability layer as an integrated part of the European research infrastructure. |  |
| URR6 | Survey, adapt, incorporate and develop ontologies and other domain conceptualizations (e.g. SKOSified authorities) as part of the infrastructure. |  |
| URR7 | Develop access mechanisms appropriate for linking across heterogeneous, potentially not curated, web-accessible information objects |  |
| URR8 | Develop knowledge management services aimed at providing better support for information, linking and collaboration between scholars and research projects. |  |
| URR9 | Develop infrastructure with a view of its use as part of a communities of practice approach. |  |
| URR10 | Develop an appropriate registry of services and tools, with metadata identifying intended communities of practice and documenting the kind and range of support provided. |  |

Table 1: User Requirement Recommendations
In order to provide an operational framework for the actual representation, analysis and understanding of evidence from scholar interviews, we adopted a cultural-historical activity theory approach, previously introduced as a useful framework in the fields of work and organization (Engeström, 2000), interaction design (Kaptelinin & Nardi, 2007), and digital curation (Dallas, 2007), and used it to develop a formal conceptual model for scholarly research activity (Benardou, Constantinopoulos, Dallas, & Gavrilis, 2010b). The model (Figure 2) can be seen as an application of the CIDOC Conceptual Reference Model, an widely accepted standard for cultural information (Crofts, Doerr, Gill, Stead, & Stiff, 2010). Entities in the DCU conceptual model for scholarly research activity, such as Actor, Research Goal, and Procedure, are specializations of CIDOC CRM entities (listed in parentheses in the model, prefixed by “E”) and are thus endowed by all properties of these entities as defined in the standard; relationships between entities, presented as arc labels in the model, are equivalent to CIDOC CRM properties or specializations thereof. Entities such as Research Activity, Procedure, Method, Information Object, Proposition, Research Goal, and Tool/Service, and properties connecting these entities, correspond to notions relevant for the conceptualization of scholarly research process by humanists, such as those sought by our empirical study presented here.

The DCU model of scholarly research activity is intended to facilitate the elicitation of requirements, and the design and development of information repositories and services in digital infrastructures that support research in the arts and humanities. For this purpose, the model does not just represent actual-historical information on a structured set of events (what, where, when was done?), but also encompasses notions of subject (who did it?), method (how was it done?) and purpose (why was it done?). This necessitated a refinement to the CIDOC CRM Activity entity, so as to introduce a distinction between Process and Procedure while maintaining corresponding (though not necessarily isomorphic) descriptions of the two; this is a duality often encountered in conceptual models of task-oriented systems, such as enterprise information systems (Dietz, 2006; Hay, 1996).

In February 2011 DARIAH completed its preparatory phase and entered the transition phase. The construction phase will commence in January 2012, after the successful establishment of a European Research Infrastructure Consortium (ERIC).

**Understanding information requirements of Holocaust scholars**

The European Holocaust Research Infrastructure (EHRI) is a four-year project (2010-2014) aimed at supporting the European Holocaust research community by establishing unified online access to dispersed sources relating to the Holocaust all over Europe and Israel, and by encouraging collaborative research through the development of relevant tools.

Following the work conducted by several organizations throughout Europe and Israel in collecting and saving fragmented and dispersed documents, objects, photos, film and art related to the Holocaust, EHRI will enable bringing all these sources together and taking the research into this area several steps further. To this end, EHRI plans to design and implement a portal, search engine and Virtual Research Environment (VRE), offering online access to a wide variety of dispersed Holocaust archives and to a number of tools to work with them.

As part of the User Requirements work within EHRI, the DCU adopted a mixed methods research protocol, intended to take forward the findings of prior research including the outcomes of our DARIAH scholarly information activity and user requirements work, through a sequential execution of a) a qualitative stage, based on semi-structured interviews, and, b) a subsequent quantitative stage, based on an extensively distributed online questionnaire (Figure 3). The first, qualitative, stage consists of the design, conduct, segmentation, coding and interpretation of semi-structured interviews with a selection of Holocaust scholars representing diverse disciplines, re-

![Figure 2: Scholarly research activity model](image)

![Figure 3. EHRI user requirements research plan workflow](image)
search topics and methodological perspectives, based on a thematic structure of about a dozen themes. Interviewing and analysis of the interview transcripts was intended to establish better understanding not only of the traditional and digital information practices of interviewees as they conduct Holocaust research, but also of their perceptions of digital research infrastructures, their role, potential affordances and usefulness. Consequently, the focus of the interviews, as well as the selection of interviewees, was planned to be purposive. In particular, it was directed to issues already known to form significant aspects of scholarly information activity (as developed by earlier scholarship, enriched by the findings of our DARIAH research, and represented in the DARIAH scholarly activity model), and that are of interest as potential sources of user requirements for the planned EHRI digital research infrastructure.

**EHRI USER QUESTIONNAIRE (excluding sub-questions)**

1. Which kinds of unpublished materials do you use in your research, in analogue and/or digital form, or both?
2. Which kinds of published materials do you use in your research, in analogue and/or digital form?
3. Rate the level of importance of each of the following activities you undertake as you seek relevant information for your research.
4. Rate the level of importance of using the following initial access points, or attributes, as you seek relevant information for your research.
5. Rate the level of importance of each of the following activities you undertake as you work with resources in foreign languages.
6. Rate the level of importance of each of the following activities you undertake as you organize unpublished materials.
7. Rate the level of importance of each of the following activities you undertake as you organize published materials.
8. Rate the level of importance of each of the following activities you undertake as you study and annotate information relevant to your research.
9. Rate the level of importance of each of the following activities as you work with others for your research.
10. Specify if you agree with each of the following statements [various normative/evaluative statements].
11. Indicate how frequently you work for your research in the following locations.
12. Indicate how frequently you use the following [digital] devices for your research.
13. For which purposes do you use computers in your research, if at all?
14. Which [software] applications do you use to store and manage research resources, if any?
15. Which of the following online services do you use for your research?
16. How would you describe yourself [in terms of professional identity]?
17. What are your main fields of research [or discipline]?
18. How would you evaluate your expertise in conducting archival research?
19. Where [i.e. in which country] do you live?

Table 2. EHRI user requirements questionnaire

The qualitative stage of our EHRI research is considered to be exploratory, as its purpose was the identification, refinement and formalisation of specific research questions, or aspects of information use by Holocaust scholars as they conduct their research. It is planned that hypotheses related to these specific questions will then be confirmed quantitatively in the second, quantitative, stage of the research. In that sense, the qualitative stage of our interview research corresponds to a context of discovery, while the qualitative stage of questionnaire survey to a a context of confirmation.

The EHRI user requirements questionnaire, now in draft, consists of nineteen questions, including demographic ones (Table 2), each composed by a further number of sub-questions, corresponding to variables at the ordinal or nominal scale. We should recall, at this point, that the thematic repertoire of the semi-structured interviews that formed the earlier, qualitative stage of our EHRI work, was elicited from the outcomes of our DARIAH research, so that it probes issues related to particular entities in the scholarly research activity model (Figure 2) we had developed as a conceptual framework for our analysis. Consequently the conceptual structure of the questionnaire (Table 2) was formulated on the basis of the thematic repertoire and the initial analysis of scholar interviews conducted up so far, and, thus specific questions bear a relation of correspondence with particular entities in our formal model of scholarly research activity: Information Object, Format, and Resource Type (questions 1-2), Research Activity and Procedure (questions 3-9), Research Goal (question 13), Tool/Service (questions 12 and 14-15), and Actor (questions 16-19). Some other questions relate to aspects of epistemic, socio-technical and physical context not explicitly covered by our model. We believe, nevertheless, that the binding of particular codes and statistical measures derived from the EHRI questionnaire survey with the structure of a well-formed and expressive ontology, such as that represented by our model of scholarly research activity, provides for a rigorous methodological framework necessary to ensure the sound interpretation of particular findings we expect to reach after EHRI user requirements research is concluded.

**Conclusions**

We present here ongoing research by the Digital Curation Unit on scholarly information work and user requirements for digital research infrastructures in the arts and humanities, as performed in the context of the Preparing DARIAH and EHRI EU-funded projects. Both projects focus to a considerable extent on developing appropriate digital infrastructure specifications to enable the effective use of text-based sources and the successful deployment of text-centred, philological and hermeneutic scholarly work. Substantive findings of our DARIAH work include, among others, the importance attributed by scholars on information curation activities, such as annotation, in tandem with information seeking activities; the complexity, heterogeneity and dense interconnectedness of information resources needed for scholarly work, bridging primary materials with a variety of epistemic objects; and, the applicability of a scholarly research activity model, developed by DCU, in actual research practices revealed through an analysis of scholar interviews.
We also advance a rationale for the mixed methods protocol adopted for our EHRi work on user requirements for Holocaust research, combining a process of discovery represented by a qualitative stage of semi-structured scholar interviews currently under way, and a process of confirmation represented by a quantitative analysis of a planned online questionnaire survey. Finally, we note the potential methodological and interpretative advantages of using a formal model of scholarly information work, based on activity theory, as a conceptual framework for the formulation of the thematic repertoire of semi-structured interviews and the identification of question included in the EHRi questionnaire.

We expect that completion of our user requirements work in EHRi, due March 2012, will allow us to develop a fuller understanding of text-based research work and a more dependable set of criteria for the evaluation of priorities in digital research infrastructures for Holocaust research. Through comparative evaluation of these results vis-à-vis earlier work in DARIAH, as well as the burgeoning body of literature on user requirements in other contemporary digital scholarly infrastructure projects, we may also be able to explore the degree to which EHRi user requirements findings are amenable to generalisation, or privy to particular fields, disciplines, or traditions of humanities research.

Acknowledgements

Authors express heartfelt thanks to all their colleagues in the DARIAH and EHRi projects, especially Sheila Anderson, Andreas Aschenbrenner, Tobias Blanm, Dimitris Gavrilis, Petri Links and Reto Speck, for useful comments and advice. They are also thankful to three anonymous SDH2011 reviewers for their constructive criticism. Finally, they gratefully acknowledge the support offered by the European Commission, the Greek General Secretariat for Research and Technology, the Athena Research Centre, and the DARIAH and EHRi consortia to research work reported in this paper.

References


Palmer, C. L. (2010). Collections in the Age of E-Research; Realizing Potential through Curation and Aggregation.


