DRAFT VERSION, submitted to DSH; published version on the publisher website, here: https://academic.oup.com/dsh/article/30/suppl_1/i43/365238

DASCH: Data and Service Center for the Humanities

Lukas Rosenthaler, Peter Fornaro and Claire Clivaz

Digital Humanities Lab, University of Basel, Ladhul, Faculty of Social and Political Sciences, University of Lausanne

Lukas.Rosenthaler@unibas.ch, Peter.Fornaro@unibas.ch, Claire.Clivaz@unil.ch

Abstract

Research data in the humanities needs to be sustainable and access to digital resources must be possible over a long period. Only if these prerequisites are fulfilled can research data be used as a source for other projects. In addition, reliability is a fundamental requirement so that digital sources can be cited, reused, and quoted. To address this problem we present our solution: the Data and Service Center for the Humanities (DASCH) located in Switzerland. The centralized infrastructure is based on flexible and extendable software that is in turn reliant on modern technologies. Such an approach allows for the straightforward migration of existing research project databases with limited lifespans in the humanities. We will demonstrate the basic concepts behind this proposed solution and our first experiences in the application thereof.

1. Introduction

The long-term preservation of digital data and resources has been an ongoing topic within the IT-industry and archiving community [i]. It is not only an IT challenge but also a challenge to both Social and Human Sciences, and academics in general, in terms of changing research and training habits in the scholarly world of the humanities. This short article focuses mainly on the former issue but will offer conclusive remarks, from a Swiss perspective, on the latter aforementioned issue.

While the Open Archival Information System (OAIS) reference model offers a very reasonable framework for the long-term archiving of digital data such as digitized motion picture, images, sound

or text documents, the archiving of highly structured digital data, such as databases, still raises a lot of problems. The flattening of databases into XML text files has been successfully used to archive the contents of relational database management systems (RDBMS) [ii] [iii] [iv]. However this method reduces accessibility since the XML files usually have to be read back into a RDBMS. In addition to this process, the logic of the application has to be reconstructed in order to regain the full usability of the data. Currently, the best method for maintaining the sustainability, functionality and usability of structured data is to migrate data repositories and their software environments (user interfaces, analytical tools and so forth) to new technologies, thus ensuring their ongoing functional accessibility[v]. The replacement of obsolete hardware and software infrastructure is an ongoing and labor-intensive process that requires continuous financial effort[v]. Furthermore, given that online research data is often being constantly modified to reflect new insights and is thus changing dynamically, referencing it (for citations as an example) is not straightforward.

Since research is not a standardized process, different research projects tend to make use of project-specific solutions and technologies. This variety of technical solutions and methods therefore leads to generic problems with respect to the interoperability and longevity of research data. Despite these difficulties, the use of digital research data, including databases, has become common in the humanities. Simultaneously, the terms used to describe the complexity of digital resources that are used and produced by the humanist researchers may be misleading: not only with respect to collections of digital objects such as digitized manuscripts; collections of digitized photographs; and metadata related to it, but also the enrichment of such sources with elements such as annotations and semantic rich links, amongst others, all of which form a base of research data in the humanities. As long as project funding is available, many of these digital sources are accessible to the research community online. However, once the funding ceases, most of these sources will remain accessible only as long as the supporting hardware and software can be kept in working order. After a certain amount of time – measurable in years – there is a danger that most of these research databases might be abandoned due to lack of maintenance. Thus, most of the digital resources and results created within research projects will have a relatively short lifespan and will no longer be available to the research community; the citation of the affected digital objects will become impossible. However, these digital sources can provide a valuable reference base for future projects but only if the maintenance thereof is ensured through continuous funding. Unfortunately such continuous financial support cannot be realized because of the

heterogeneity and lack of critical mass of most project-related databases.

Furthermore, many of the printed publications (paper or e-paper such as PDF[vi]) published during the lifespan of a research project rely on results obtained from the original digital data collections. In order to add the often-demanded transparency (traceability and conformability), the source material that the research is based on should also be available for critical review.

2. Approach

Given this challenging situation, the Swiss Academy of Humanities and Social Sciences (SAHSS) - on the behalf of the State Secretariat for Education, Research and Innovation (SERI) - has launched a project to address this situation in the national context of Switzerland. The Digital Humanities Lab of the University of Basel (DHLab) in conjunction with the Universities of Lausanne and Bern, in association with the Swiss National Archives, participated in a call for proposals resulting in the joint departments being given the task of establishing a solution. In the initial three-year period (2013-2016) is a pilot program leading to the establishment of a Data and Service Center for the Humanities (DASCH). This program will be based on several test cases of different sizes and complexities from different disciplines; the methods and processes, legal aspects, infrastructure needs, and finally, the costs and expenses, all have to be evaluated. The proposed DASCH is based on the following premises:

A) Preserving software in a usable and working condition remains a difficult task, as illustrated at the recent meeting 'Preserving.exe: Toward a National Strategy for Preserving Software' held by the American Library of Congress [vii]. It would therefore be too difficult and costly to maintain a multitude of different systems for a long time.

B) Emulation of obsolete hard- and soft-ware as proposed by R.A. Lorie (Lorie, 2001) is also not an easy approach, having its share of problems (Luan *et al*., 2010). Instead, with respect to DASCH, we propose that only a minimal number of centralized service locations should be operated: where the different digital sources or databases will be integrated through an importation and translation process. As a result, ideally only one type of technological infrastructure has to be maintained. DASCH as defined by the call for proposals addresses the following:

- The first phase focuses on the adoption of existing data resources of research projects that have reached the end of their funding. In the second phase, the goal is for researchers of ongoing or future research projects to be guided through the creation and use of digital sources

in order to facilitate the accessibility of the data once funding has ended.DASCH facilitates the exchange between platforms and infrastructures by implementing interfaces for the import, export and querying of information (data restricted by means of legal constraints, such as copyright issues and/or protection of personal rights, are mapped in a sophisticated access right system). The adopted digital sources must be accessible through a flexible and extendable user interface that allows for search and analysis. We are currently supporting a RESTful (Representational State Transfer, a standard for scalable web services) web-service and a Linked Open Data SPARQL-Endpoint in order to integrate the data into other research projects and/or databases.DASCH should encourage new collaborative research models in order to allow for the optimal use of digital sources. It should also be able to facilitate efficient training modules and support for all the new research projects funded by the Swiss National Science Foundation. In order to prepare for the interrelation of Swiss digital Humanities research to international research, international contacts are a key focus for this center.

Given the nature of research in the humanities, the data sources DASCH has to deal with are heterogeneous and consist mainly of qualitative data (which is possibly linked to digitized objects). The experience thus far, with respect to the integration of approximately fifteen projects of different sizes and complexities, has shown that anything from simple spreadsheets to complex, undocumented relational databases with more than a hundred tables can be expected. The SALSAH-platform[viii] was chosen as base for the consolidation of different data sources. The repository of SALSAH is based on semantic web technologies (RDF, RDFS and OWL), and it has a modular, layered architecture and is thus well suited to the emulation of the basic functionality of RDBMS's, simple databases such as MS-Access, and other databases such as FileMaker. SALSAH is currently being actively developed by the DHLab and will be available as open source by the end of 2015.

=> figure1.tif

Fig. 1 The web interface of SALSAH implements a desktop interface within the web-browser window in order to allow for the visualization of a working RDF-graph with multiple sources simultaneously.

=> figure2.tif

Fig. 2 SALSAH allows a dynamic visualization of the graph-like structure of the RDF data representation.

=> figure3.tif

Fig. 3 The architecture of the software infrastructure is highly modular, flexible and extendable; the design follows strict layering.

Within a research project funded by the Swiss National Science Foundation, SALSAH is currently being extended with several new important features (expected in 2015)

- A time machine that will allow digital objects to be referenced by permalinks which include a timestamp for referencing. Such permalinks will always display the digital object in the state it was in at the time it was referenced. These permalinks will add true 'citation-ability' to the SALSAH environment.
- SALSAH, which is currently organized technically as a centralized system, will be transformed into a distributed, self-organizing P2P system. At the same time, an archival system based on DISTARNET (DISTributed Archival NETwork, a self-organizing peer-to-peer network for archival storage of digital data, see Subotic *et al*., 2013) will be added to SALSAH for protection against data loss due to catastrophic events like hardware failure, flooding, and fires occurring at any SALSAH location. DISTARENT also uses P2P technology to maintain redundant multiple backups of the data within the network.
- Within the DASCH project, SALSAH will be extended to support open data-standards [11] for access and linked data [12] . However, open access may be restricted by legal reasons (such as copyright and privacy laws). SALSAH includes fine-grained identity and rights management.
- SALSAH will be continuously enhanced according to the needs of the researchers using the platform. It is planned for SALSAH to migrate to open source by the end of 2015.

The main tasks of DASCH will be threefold:

1. Maintaining the technological infrastructure and adapting it to the needs of researchers and changing technology. This task will be handled in Basel during the pilot phase but since SALSAH will become an open source project, other institutions and individuals may contribute to its base. However, in our experience, open source projects need a powerful coordinating institution in order to be successful. The DHLab will be on hand to play this role.
2. Assistance to researchers: in the humanities, many researchers working with digital sources do not have the technical knowledge to fully exploit the advantages of digital processes. DASCH will support researchers, with regards to the best possible use of digital methods and tools for

their research. Training and education will also be encouraged.

3. To create a report with recommendations with respect to proceeding with the project and transferring the pilot project to a permanent institution.

It should be noted that funding for this project goes far beyond normal scientific project financing, this is evidence of the strong commitment of all parties involved (SAHSS, SERI, Swiss National Science Foundation) in the foundation of a sustainable national data curation and service center for digital research data in the humanities. The team comprising of the Universities of Basel, Bern and Lausanne, and the National Archives demonstrates this commitment. The project's pilot phase is financed by SAHSS and SERI. Since Switzerland is a multilingual nation with a highly federalized structure, we decided to create DASCH as a virtual center where the technological infrastructure will be located and maintained in Basel for the time being, but all the other tasks will be performed by local branch offices that are very close to the researchers. During the pilot phase, Lausanne and Bern will be testing this branch-office, satellite model. As soon as SALSAH's P2P functionality is implemented, the technical infrastructure may be distributed as necessary or desired. Due to the DISTARNET archival system, data is secured against loss without the local branches needing to build an expensive and complicated backup infrastructure. As a final point of focus, the IT element of this infrastructure project requires explanation with respect to the Swiss institutions. Indeed, until now, the funds devoted to research and the funds focused on infrastructure have been always separated. Through this pilot project, in collaboration with other colleagues working on DH projects, we hope to progressively overcome, at least in part, this defined division. Even the best system of data curation has to overcome a significant obstacle: the (un)willingness of a SHS researcher to give their digital material to a platform. To overcome this recurrent difficulty, several solutions have to be explored. Firstly, DH training and education is an influential element that directly influences research and IT developments. Secondly, with respect to the usual individual habits of SHS researchers, a top-down project such as this one could be complemented by a bottom-up approach, represented, for example, by the recent NAKALA platform created by Huma-Num[ix]. Indeed, to promote SALSAH as a solution for all the DH material in a country would be a near impossible task in human, financial and concrete terms. We would rather, through SALSAH, encourage and stimulate data-life-cycle management in the humanities. In such a proposed center, parallel services could be developed, such as copyright counseling for images and online data.

These remarks are surely not only related to possible further developments: as the striking example of the Arts and Humanities Data Service (AHDS) in UK proves «a United Kingdom national service aiding the discovery, creation and preservation of digital resources in and for research, teaching and learning in the arts and humanities, [...] established in 1996 and ceased operation in 2008»[x]. This project however, had too great a focus on technical challenges and was unable to offer the services and research that make DH truly come alive. The overcoming of the division between research funding and infrastructure funding is surely one of the key-challenges of such regional/national projects.

---

[i] e.g. Kuny, Terry (1997). A Digital Dark Ages? Challenges in the Preservation of Electronic Information. In: 63rd IFLA General Conference: Conference Proceedings. International Federation of Library Associations and Institutions.

[ii] e.g. SIARD of the National Archives of Switzerland, http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en; last accessed on 14/10/14

[iii] XENA of the National Archives of Australia, http://xena.sourceforge.net; last accessed on 14/10/14

[iv] kopal/koLibRi of the german Kopalproject, http://kopal.langzeitarchivierung.de/index_koLibRI.php.en

[v] e.g. Rosenthaler, L., R. Gschwind und F. Frey (1999). The Digital Age in long-term Image Archival – Risks and Prospects. In: ICOM-CC.

[vi] which we consider to be equal to a printed paper from a technical point of view

[vii] Preserving.exe: Toward a National Strategy for Preserving Software, May 20-21, 2013, Library of Congress, Washington DC, see also http://www.digitalpreservation.gov/meetings/documents/othermeetings/preservingsoftware2013/Preserving_Exe_Agenda.pdf; last accessed on 14/10/14

[viii] System for Annotation and Linkage of Sources un Arts and Humanites, a virtual research platform, see http://www.salsah.org for the generic entry point, http://www.salsah.org/dokubib for an (simplified) entrypoint for the Documentation Library of St. Moritz, and http://www.salsah.org/kuhaba for the Kunsthalle Basel.

[ix] https://www.nakala.fr/; last accessed on 13/10/14. Huma-Num means “la très grande infrastructure des humanités numérique”, see http://www.huma-num.fr/

[x] See http://en.wikipedia.org/wiki/Arts_and_Humanities_Data_Service; last accessed on

13/10/14.

Subotic I., Rosenthaler L. and Schuldt H. (2013). A Distributed Archival Network for Process-Oriented Autonomic Long-Term Digital Preservation, ACM Proceedings of the Joint Conference On Digital Libraries, ACM New York (2013), pp. 29-38

Lorie R. A. (2001). Long term preservation of digital information, Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, Roanoke, Virginia, United States. 24–28 June 2001. New York, NY: Association of Computing Machinery. pp. 346-352

Luan F., Nygard M. , Mesti T. (2010). "A survey of digital preservation strategies", World Digital Libraries, Vol3 (2), IOS Press