Digital Humanities 2016
Kraków, 11–16 July

(http://dh2016.adho.org)

Title: Implementation of a National Data Center for the Humanities (DaSCH)
Authors: Lukas Rosenthaler, Beat Immenhauser, Peter Fornaro
Category: Paper:Long Paper
Keywords: semantic web, data center

Rosenthaler, L., Immenhauser, B., Fornaro, P. (2016). Implementation of a National Data Center for the Humanities (DaSCH). In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 339-341.

Implementation of a National Data Center for the Humanities (DaSCH)

1. Introduction

Up-to-date research in the humanities today depends as much on digital methods and digital data. However, the use of computer-based methods and online sources in the humanities still faces several challenges, including the difficulty of ensuring the longevity of research data, the lack of common basic services, inadequate standardisation of data formats, insufficient training in digital methods and best practices, and weak international Digital Humanities networks. Digital documents are accumulated, organised and annotated using electronic databases. However, the necessary infrastructure is most often established in a project- specific way and is not designed for the long-term preservation of data. After the completion of a research project, these digital resources quickly become unavailable if they, and the software and hardware they rely on, are not properly maintained. Keeping digital data accessible after the end of a project is costly in terms of money and labour and is usually not included in the project funding.

While the digitisation of analogue sources produces large numbers of digital documents, these documents usually have a simple structure. By contrast, the data produced during the research process is much more complex, consisting of interlinked information (databases, annotations etc.). Because of the complexity of this research data, it is very difficult to make it permanently available. However, there are several reasons for doing so:

Transparency:
As research data is the foundation on which published results are based, it be- comes necessary to have access to this data in order to evaluate the results.

Reuse:
New research projects can reuse existing research data to propose different answers to the same questions, or to ask entirely new questions, especially if the datasets from different projects can be linked.

Citability:
Digital sources may only be referenced in scientific texts if they can be accessed permanently without modification. The long-term accessibility of arbitrary digital objects (together with permanent links and unique object identifiers) is usually not possible.

2. Organisational form

The Swiss Academy of Humanites and Social Sciences (SAHSS) therefore decided to establish in collaboration with the

Digital Humanities Lab (DHLab) of the University of Basel a new “national research infrastructure” (Data- and Service Center for the Humanities, DaSCH) which takes this kind of digital research data into custody and preserves the direct online access. The primary goals are:

- Long-term curation of research data
- Permanent access and reuse
- Services for researchers to support data life-cycle management

The secondary goals are:

- Promoting the digital networking of databases created in Switzerland or in other countries
- Carrying out a pilot project in close proximity to humanities research
- Collaboration and networking with other institutions on developing digital literacy

During a pilot phase lasting two years that ended in July 2015, the data of about 25 different research projects ranging from ancient history to musicology have been passed to new institution for preserving long term accessibility. In order to copy with such heterogeneous data, the platform has to be extremely flexible and versatile.

Since Switzerland is a highly federalist country, a balance between a central/decentral approach had to be chosen. We decided to form of a network that currently consists of several “satellite” nodes and a central office which acts as coordinator, main provider of technology and software development. The individual locations have a great deal of freedom to take local decisions (e.g. which research projects are considered important to be included in the platform). At each satellite location, it is necessary to have both a broad knowledge and experience available in humanities research as well as in IT and software development skills. The central office provides second-level support.

3. Technological issues

Our daily experience seems to suggest that digital data is quite volatile and unstable. Everybody who works with computers on any scale has suffered the unfortunate experience of data loss. In a recent interview, Vincent Cerf, often regarded as one of the "fathers of the internet", says he is worried that all the images and documents we have been saving on computers will eventually be lost: "Our life, our memories, our most cherished family photographs increasingly exist as bits of information – on our hard drives or in "the cloud". But as technology moves on, they risk being lost in the wake of an accelerating digital revolution." (Cerf, 2015) Thus, it appears that "long-term archival" and "digital" are diametrically opposed concepts. However, the digital domain offers some unique characteristics that allow the long-term preservation of digital data. However, guaranteeing long-term access to digital information remains a tedious and difficult process.

There are only a few fundamental methods for long-term preservation of digital data:

Emulation
The software and to some extent the hardware of obsolete computer system can be emulated ("simulated") on modern computers. Thus data can be rendered using vintage software.

«Eternal» media
The «eternal» media approach requires the digital data to be recorded onto the most robust and durable media available.

Migration
In the context of long term archiving, migration is defined as the process of periodically copying digital data onto new, up-to-date storage media and, if required, converting the file formats to new, well-documented standard formats.

The OAIS reference model for a digital archive is based on the migration model. In addition to a formal process description, it also covers the ingest of data into the archive and the dissemination of archived data to a user. An important aspect of the OAIS reference model is the systematic approach to metadata that is distinguished between the metadata required to identify and find a «document», and the technical metadata required for the management of the migration processes. The OAIS approach can be adapted for complex «objects» such as relational databases or NoSQL-databases (e.g. using the SIARD-suite (Ohnesorge, 2015), a standard adopted by European PLANETS project and as Swiss eGovernment Standard eCH-0165), however in order to browse or use the data, the whole dataset has to be retrieved from the archive and converted back into a working RDBMS using the SIARD-Suite – a «quick overview» is not possible.

Complementary to the OAIS archival process model, *keep-alive archiving* keeps a system of data, data management and access methods online and permanently up-to-date. Whenever the technology evolves (e.g. a new stable version of the data management software or a new version of a file format is released), the whole system is migrated to conform to the new environment. The keep-alive archives are especially well suited to complex data such as databases which are accessed very frequently. However, there two fundamental problems with keep-alive archives:

If the data management system does not offer a method to record all changes, the history will be lost.

It is virtually impossible to keep each projects IT-infrastructure – especially the software – running forever. Each project uses its own software (Filemaker Version XY, MySQL, PHP, ruby, Excel, etc.) and data models. The adaption to the evolving technology would overwhelm each institution.

The DaSCH implements a modified keep-alive concept. It has chosen to use the Resource Description Framework (RDF, standardised by the W3C) as a common ground for representing the data. It provides a very simple but highly flexible representation of digital information. RDF allows the definition of ontologies which formalise the semantic relationship of digital objects. We defined a base ontology which implements some required basic concepts (e.g. timestamp based versioning, annotations, access rights etc.). Starting from this base ontology, for each research project taken into custody a specific ontology is being derived. On delivery of the data, the original data structure is translated into this ontology preserving the important features and relationships of the data. This technological framework thus allows the «simulation» of almost any data models (relational databases, XML hierarchies, TEI-encoded texts, graph networks etc.) in a common infrastructure using open standards such as RDF, RDFS [1] and OWL [2].

The pilot phase has made it clear that project-specific access applications (such as online graph- ical user interfaces) have to be preserved. While this approach does not make it possible to directly reuse the original applications, it has been shown that is easy to re-implement their basic functionality as well as their look and feel.

Using the common platform, it is straightforward to create new tools and applications that reuse existing data by combining information from different datasets. Thus, new research methods can be implemented, e.g. using methods of «big data» analysis

Due to the success of the pilot phase where about 25 projects have been integrated, some with individual user interfaces, the Academy has decided to ask for funding. The request is awaiting the approval of the swiss national parliament.

Bibliography

1. **Cerf, V.** (2015). Interview on BBC http://www.bbc.com/news/science-environment-31450389 (accessed March 4th 2015).
2. **Ohnesorge, K., Mérinat, T. and Büchler, M.** (2015). SIARD Format Version 2.0, SFA I 2015-10-15 I DLM Forum 2015, Luxembourg, http://www.eark-project.com/resources/conference-presentations/dlm-oct15/37-siard2eark-1/file (accessed March 4th 2016).

Notes

[1] RDS-Schema for expressing simple ontologies.

[2] Web Ontology Language for expressing complex ontologies and relations.