Long-term Access to Research Data as a Challenge to Migration

Tobias Schweizer, Andreas Wassmer, Ivan Subotic, Lukas Rosenthaler; Digital Humanities Lab, University of Basel; Basel, Switzerland

Abstract

Digitization allows humanities scholars to access their primary sources with the help of computers. As a consequence, research data is also created within digital infrastructure by using databases, annotation tools or virtual research environments. This novel way of using the computer in the humanities exceeds the conventional use of word processors. The dynamic nature of this data poses new problems beyond the difficulty of storing it for a long-term period (archiving): data should also be permanently accessible and usable to base future research upon.

Because of rapid changes in hardware and software, migration of research data to a working and long-term supported infrastructure is indispensable. SALSAH is developed as such an infrastructure. It is a web-based generic research platform for the humanities allowing for the collaborative annotation and linking of digital sources.

Introduction

After years of digitization, huge amounts of physical sources (manuscripts, books, photographic material, etc.) have been made available in a digital form for humanities research. Unlike the physical sources themselves, their digital representations can only be accessed by using a computer, thereby changing the way research is carried out in the humanities. The computer can no longer be seen as a 'better typewritter', but as an *infrastructure* for research. Databases, annotation tools and virtual research environments are being used more and more to work with digital sources in the humanities.

In natural sciences there is a long "tradition" in generating digital research data and researchers are aware of the importance of keeping this data accessible. Jim Gray [3] even saw a fourth paradigm emerge. The first paradigm, the experiment, is the oldest one, dating back to the 17^{th} century when systematic research began. In the 20^{th} century, the development of theoretical models (the second paradigm) became more and more important which was then accompanied by methods depending heavily on computing power (the third paradigm). This last change of paradigm generated a huge amount of data because mostly experiment, theory and results generated by software are joined to form the new research data. It is the exploration of this data that Jim Gray identifies as the fourth paradigm. This demands a reliable data curation infrastructure which has yet to be built.

As in science, research data in the humanities today is very likely to be created, processed and stored digitally. In contrast, the product of humanities research can still (and often still has to) be a conventional monograph or an article in a printed journal. Books and journal articles – the products of research – can easily be digitally simulated by creating PDFs or websites, and e-publishing offers added value like search functions or hyperlinks, etc.

Archives, both conventional and digital, take care of preserving these documents and providing access to them, as well as metadata and retrieval procedures to these documents. But when it comes to research data, this archival approach is *insufficient*. Unlike the finalized documents, research data has to remain both functional and extensible so that future research can be based on it. Without this precondition Johannes Kepler would not have discovered the laws of planetary motion. He was able to deduce these laws by carefully studying Tycho Brahe's catalog of systematic astronomical observations.

Quite recently, the German company "PediaPress" announced a crowdfunding campaign to print the whole English Wikipedia in more than a thousand volumes [1]. Printing Wikipedia seems to be a prestige project to show how extensive the on-line encyclopedia has become. But it holds also some problems. The most obvious one is the fast obsolescence of printed articles since the on-line Wikipedia is continually updated and changed. But this difficulty is well known in the print world and can be solved by providing new editions. A more severe problem is that a printed version of Wikipedia and also a mere PDF-version lack the basic *functionality* which makes Wikipedia what it is: almost every article can be instantly updated by anyone, previous versions of articles remain available and can be cited as such, and the discussions among editors in the revision process are also accessible. Printing Wikipedia articles means finalizing them: the process (or so to speak: the research data) that led to the article in its current form is excluded from the article's presentation.

In this article, we will introduce SALSAH, a virtual research environment for humanities disciplines, and discuss its relation to common digital archiving concepts. We will present our concept of a long-term supported research infrastructure and its data model, which is able to integrate research data from other environments in order to prevent its technical obsolescence. We will illustrate our explanations by providing two use-cases.

Related Work

The Library of the Swiss Federal Institute of Technology, Zurich, provides a digital curation service to its scholars [2]. The service ensures preservation of all file formats in their original condition. Data uploaded to their infrastructure can be shared within the research group but also with third party researchers. If necessary, the data can be published world-wide by distributing a unique digital object identifier (DOI). This can also be used to permanently reference research data in publications.

The University of Minnesota conducted a data curation pilot in 2013 [5]. In their project they developed and implemented a curation workflow for 5 different examples of research data. The examples ranged from engineering data to health sciences and social and humanities data. The data came in many file formats such as Excel, video, images and as paper prints. In their final report they listed 4 lessons learned regarding the curation of research data: (1) In some cases faculties were not too much concerned with sharing their data with others but more with finding a permanent home for them as a way to back them up, (2) it is very important to acquire an expertise in most of the file formats and in the use of the applications that generate them, (3) there is a need to educate the researchers regarding the documentation of their datasets, and (4) although all datasets selected for the pilot were considered to be ready for public access and reuse, there were ownership and intellectual property issues regarding nearly all the data.

In chemistry a lot of research data is generated by machines, e.g. mass spectra from a mass spectrometer. The spectra which are rich machine-readable data are saved to a storage media and almost always transformed to human-readable images or PDFs. This process often renders it almost totally un-indexed and undiscoverable. Furthermore, the raw data often comes in a proprietary format linked to a certain analyzing device. Publishing this data is therefore almost useless to other researchers. The project "Chempound" tries to overcome these difficulties [4]. It is a repository architecture based on RDF. Chempound is made of two components: an RDF store to describe the structure of the data held in Chempound and a resource store which holds the research data. But legacy file formats are not simply stored in a database. They are imported as a package of files with associated metadata. On import, the content of the package is further translated into Chemical Markup Language (CML, an XML dialect) and RDF data is added. The conversion to CML renders the data into a common data structure which can be used for the generation of domain-specific output data if necessary.

There are many more related projects going on. The curation of research data has become a very active field of research. To us, an approach based on the RDF model like in the Chempound project seems very promising. This allows for the setup of a flexible data curation system, that can import and link a huge variety of data and keeps their semantics at the same time. Nevertheless, there is no general product available yet. That motivates the further development of our research infrastructure for humanities disciplines.

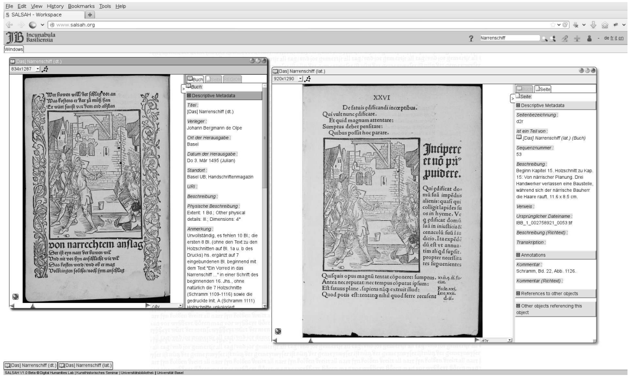
SALSAH

SALSAH (**S**ystem for **A**nnotation and **L**inkage of **S**ources in **A**rts and **H**umanities) is a virtual research environment for humanist researchers that has been developed at the Digital Humanities Lab at the University of Basel since 2009. SALSAH started as a subcomponent of an art historical research project on image-text relations in Sebastian Brant's 'Narrenschiff' ('Ship of Fools'), that is held in the Basel University Library and had been digitized along with other incunabula (see Figure 1). Originally designed to present the results of art historical research on the web, SALSAH was soon conceived to be an apt tool to carry out research with. SALSAH is seen as a 'tool-provider' providing generic instruments for annotating and linking digital resources. In the past four years, research projects from various humanities disciplines have become part of SALSAH such as musicology, the study of literature, cultural anthropology, etc.

Functionality

SALSAH is a completely web-based research environment. That means, that all of its (client-side) functionality runs in a web-browser[1]. SALSAH's workspace is implemented with HTML(5)/CSS and JavaScript code based on the jQuery-framework (`http://www.jquery.com`), without the need to install any third-party addons.

SALSAH simultaneously visualizes various digital objects by means of a window system: each digital object is represented by a window in the SALSAH-workspace (within a *single* browser window or tab) which can be rearranged, resized and minimized. A digital object can be an abstract entity or have a concrete digital representation such as an image, text, video or audio file. SALSAH's functionality is generic: every digital object can be annotated (related to a semantically defined value) and linked to other digital objects. The links themselves can be annotated as digital objects in order to express their meaning.



Buch:

Descriptive Metadata

Titel : [Das] Narrenschiff (dt.)

Verleger : Johann Bergmann de Olpe

Ort der Herausgabe : Basel

Datum der Herausgabe : Do 3. Mär 1495 (Julian)

Standort : Basel UB; Handschriftenmagazin

URI :

Beschreibung :

Physische Beschreibung : Extent: 1 Bd.; Other physical details: Ill.; Dimensions: 4°

Anmerkung : Unvollständig, es fehlen 10 Bl.; die ersten 8 Bl. (ohne den Text zu den Holzschnitten auf Bl. 1a u. b des Drucks) hs. ergänzt auf 7 eingebundenen Bl. beginnend mit dem Text "Ein Vorred in das Narrenschiff ..." in einer Schrift des beginnenden 16. Jhs., ohne natürlich die 7 Holzschnitte (Schramm 1109-1116) sowie die gedruckte Init. A (Schramm 1111) Holzschnitte unkoloriert

Seite:

Descriptive Metadata

Seitenbezeichnung : d2r

ist ein Teil von : [Das] Narrenschiff (lat.) (Buch)

Sequenznummer : 53

Beschreibung : Beginn Kapitel 15. Holzschnitt zu Kap. 15: Von närrischer Planung. Drei Handwerker verlassen eine Baustelle, während sich der närrische Bauherr die Haare rauft. 11.6 x 8.5 cm.

Verweis :

Ursprünglicher Dateiname : IBB_1_002758921_0053.tif

Beschreibung (Richtext) :

Transkription :

Annotations

Kommentar : Schramm, Bd. 22, Abb. 1126.

Kommentar (Richtext) :

References to other objects

Other objects referencing this object

***Figure 1.** Sebastian Brant's 'Narrenschiff' (Basel UB, Ai II 22b:1, c4v) and its Latin Translation 'Stultifera navis' (Basel UB, DA III 1, d2r) in the SALSAH Workspace*

Figure 1 shows the two incunabula the 'Narrenschiff' and its Latin translation in SALSAH's generic compound viewer[2]. The single pages can be browsed by using a slider, and the associated metadata and annotations are presented aside. Given the necessary access rights by logging in, a researcher may change annotations or create new ones together with other persons (collaboration). For images, SALSAH provides the region-of-interest functionality (ROI). A researcher may define geometrical shapes (rectangles, circles or polygons) on an image and annotate these regions as distinct digital objects. The regions are not really drawn on the image, but overlaid with the HTML5 canvas-element using its 2D-drawing context (JavaScript). Based on the ROI-functionality, manuscripts can be addressed topographically and transcribed. For video and audio data, a transcription tool is being developed allowing for the definition and annotation of sequences.

Summing up, SALSAH is a complex web application providing both access to sources[3] and appropriate tools for creating annotations and annotated links among them. The researchers always deal with the SALSAH user interface and do not have to care about where or how their data is stored. Of course, SALSAH

[1] SALSAH works with the current versions of Firefox, Chrome and Safari.

[2] A compound object is a digital object that consists of other digital objects like a book that contains pages or a photographic album containing photographs.

[3] The sources do not have to be necessarily stored inside SALSAH, but can also be referenced from digital repositories (see section Linking of Digital Repositories to SALSAH).

offers ways of exporting research data into portable formats such as XML, PDF, etc.

Research Data and Digital Archiving

In discussions of research data in the humanities, the term 'data curation' has emerged. As the authors of an introductory article argue, 'curation' as known from the context of museology gets an additional meaning here:

> The curation of research data - raw and abstracted material created as part of research processes and which may be used again as the input to further research - carries with it the burden of capturing and preserving not only the data itself, but information about the methods by which it was produced. If the methods used to generate the data are algorithmic, the method itself may need to be captured and curated. Because these methods and information on the goals involved in creating the data are often essential to its subsequent interpretation and reuse, they can be considered an important part of the data itself. In addition, because reuse is such a crucial aim, successfully curated data needs to remain functional, and this may require regular changes to its state. [6]

The way we understand the cited passage, the focus lies on research data as documenting the process of research and remaining in a non finalized state that allows for further research. Unlike citing an article or a monograph (finalized results of research) and writing *about* them, being able to directly access research data in a functional form means carrying on with research conducted by someone else with his or her *own* means (the way he or she used to work – the tools, methods, etc.). Maintaining and giving access to research data in the humanities is a new concept because it goes beyond the conventional publishing of research results, cf. [7].

This paradigm is *not* compatible with the concept of digital archiving or preserving. The OAIS reference model defines archiving as maintaining information over an unspecified period of time. To this end, the data itself – the bitstream – has to be preserved along with a documentation – the representation information – how it has to be interpreted in order to arrive at the original information at access time [8]. That means that between the submission of information to a digital archive and accessing it at a later point in the time, it is not in a directly usable or functional state. The aim of a digital archive is to preserve information by regularly copying the data and – if necessary – migrating it to other formats to ensure its interpretability, whereas we are trying to maintain the research data in a *permanently* usable and functional state. This implies a permanent migration of data formats and the continuous maintenance of the research application.

Migrating Research Data

Given the increasing usage of complex software infrastructures in the humanities research, their maintenance in the long-term becomes a problem. Generally, research projects in the humanities are limited to a few years due to funding policies. E.g., in Switzerland project funding usually covers three to four years. Often, a digital research infrastructure built up during a project can no longer be supported after the project's end because the responsible persons are no longer available or the technologies used have become obsolete. In order to guarantee further access to research data in a functional state, a *migration* of the research data to a working and long-term supported infrastructure is indispensable.

With SALSAH, we are maintaining a central infrastructure for carrying out research in the humanities in Switzerland. We are assuring the funding for SALSAH on a national and long-term basis. SALSAH's data model is very flexible so that it can integrate other projects' digital infrastructure: e.g, the data model and data of a relational database can be transfered to SALSAH in order to guarantee its further usability.

SALSAH's Data Model

SALSAH's data model is based upon the Resource Description Framework (RDF) and RDF Schema (RDFS). RDF is a W3C recommended standard for building the semantic web [9, 10]. RDF allows the representation of information about resources by forming triples: a *subject* (resource) is assigned a *predicate* (property) with a certain *object* (value).[4] The subjects and predicates are identified by Uniform Resource Identifiers (URI). An object can be a concrete value (literal) or another subject identified by an URI. In this way, resources can be described (e.g., by assigning information conforming to the Dublin Core metadata schema like 'author' or 'title'), but also *linked* to other resources. The result is a network that can be semantically described: by following the principles of RDFS, resource and property vocabularies (namespaces) can be defined. Each resource and each property refers to a semantic concept.

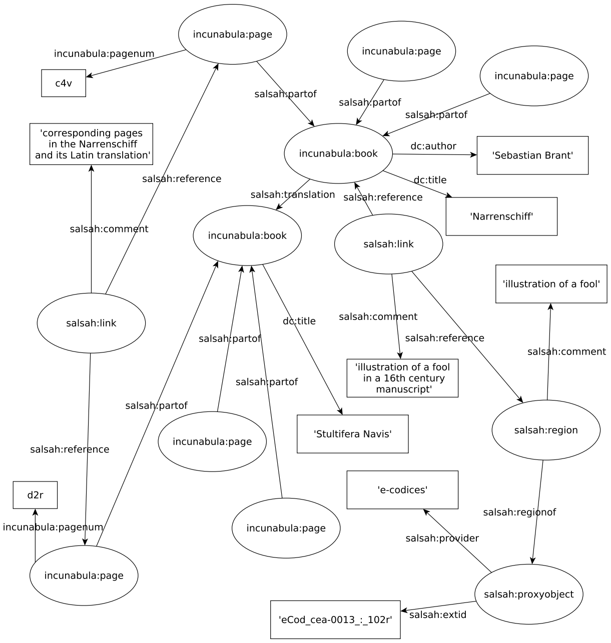


Figure 2. *Illustration of SALSAH's data model*

Figure 2 shows some RDF-triples contained in SALSAH.[5]

[4] This is the official RDF-terminology. However, we will use the terms subject/resource, predicate/property. and object/value interchangeably.

[5] For reasons of clarity, only a few of the existing triples are represented.

The resources are represented by ellipses, the properties by edges and the values by rectangles (unless they are resources themselves). Figure 2 shows the triples underlying the front-end presentation shown in Figure 1. Both Sebastian Brant's 'Narrenschiff' and its Latin translation are instances of the resource-class 'book' defined in the 'incunabula'-vocabulary. These resources can be described by properties defined in the 'incunabula' vocabulary, but also by using 'generic' properties from the 'salsah:vocabulary': whereas 'incunabula' is a project-specific vocabulary, the 'salsah'-vocabulary can be used comprehensively. Figure 2 shows the corresponding pages of the two books, and this relation can be made explicit by making use of the 'link' resource. The link-resource connects two pages and provides information why they are related by providing a 'comment'.

They bottom right part of Figure 2 shows another 'link'-resource. But unlike in the case described above, here a SALSAH-resource is related to a 'region'-resource belonging to an *external* resource represented by a 'proxyobject'-resource. The 'proxyobject'-resource allows for the annotation and linking of external resources as *if the were locally contained in SALSAH*. By supplying information about the provider and the provider-specific identification, the local annotations can be mapped to the information hosted at a remote location (see section Linking of Digital Repositories to SALSAH for more details).

Migration of a Relational Database to SALSAH

In this section we will provide some information how an existing relational database can be transferred to SALSAH in order to guarantee its further usability and functionality irrespective of the future condition of the environment it has been created in.

As an example for demonstration purposes, we haven chosen the 'Kritische Robert Walser-Ausgabe' which edits all printed texts and manuscripts of the Swiss writer Robert Walser [11]. The edition is realized as a series of printed books accompanied by a digital version on DVD. For the preparation of the edition, a MySQL database is used. This database, or rather a part of it, will be transferred into SALSAH and held in a synchronized state there (using SALSAH's RESTful API): when metadata change or are newly created, they will be 'pushed' into SALSAH.[6] In a current test stage, SALSAH is being used to present digital facsimiles (photographic reproductions of documents) along with the corresponding metadata created in the MySQL-database. It will then be possible to annotate and link the digital facsimiles in SALSAH. For legal reasons, the data and the digital facsimiles will only be accessible with the necessary rights (by logging in).

Figure 3 shows the simplified[7] data model of the 'Kritische Robert Walser-Ausgabe' in MySQL. It represents the arrangement the edition is based upon. In this example, the materials consist of journals and newspapers Robert Walser has published in.[8] A 'document' is a certain issue of a newspaper or journal, that can be identified by a title and bibliographical information. It is also characterized by a type (whether it is a journal or a newspaper).

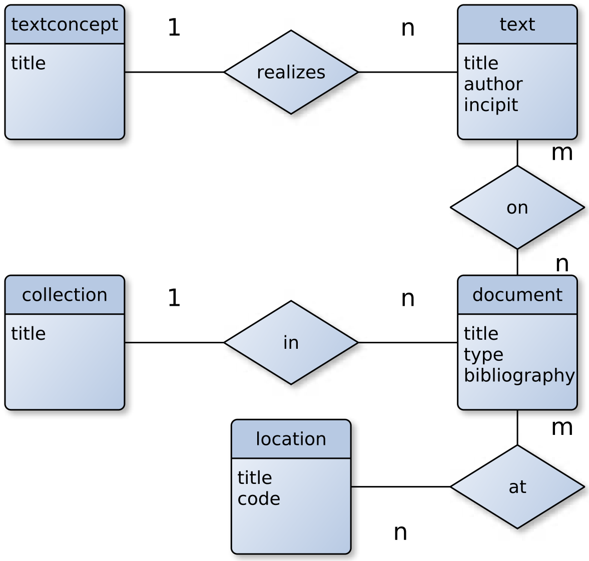


Figure 3. *The Data Model of the 'Kritische Robert Walser-Ausgabe' in MySQL*

A 'document' carries 'text', which can be described by a title, the incipit (the first sentences) and the author (mostly Robert Walser himself, but possibly also reviewers of his texts). While a 'text' is concrete (it has been printed and can be read), the 'textconcept' is abstract: it is an intellectual concept and can be related to multiple 'texts'. Some 'texts' might be similar and can be conceived to be variants respectively different readings of the same 'textconcept'. A 'document' can be held at various locations (archives, libraries) and several 'documents' can be grouped to a 'collection'.

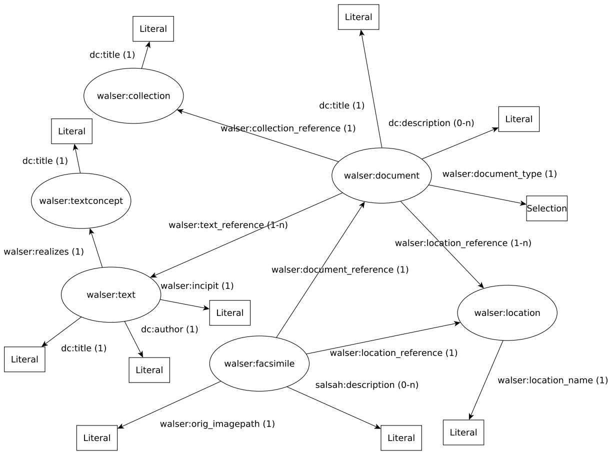


Figure 4. *The Data Model of the 'Kritische Robert Walser-Ausgabe' in SALSAH*

In order to transfer the described data model into SALSAH, it has to be adequately expressed using RDF(S). By saying adequate, we do not mean a mechanistic transformation of a relational data model to RDF. Rather, we want to primarily meet the requirements of the domain. In consequence, the resulting data model in SALSAH is not a direct mapping of the relational one

[6] That means, that the MySQL-database and SALSAH run simultaneously, but MySQL remains the master database for the metadata.

[7] The data model in MySQL is far more complex and extensive than shown in Figure 3.

[8] Of course, materials could also be manuscripts or the famous 'Mikrogramme'.

since we do not want to reproduce characteristics of a relational data base. But of course, the relational data model has to translatable into the new data model by an import process.

Figure 4 illustrates the data model in SALSAH. The information in parentheses on the edges reflects the cardinality of a property: 0-1, 0-n, 1 or 1-n. The RDF model is simpler than the relational one, in that no auxiliary tables are required (as present in m:n relations). After the import the full functionality of SALSAH can be used on this data. In this data model, also the digital facsimiles are present. They represent documents (a certain issue of a newspaper or a journal, not a concrete copy of them) and are linked to their provenance (the archive where they have been produced).

Linking of Digital Repositories to SALSAH

In Figure 2 we have also presented a resource of the type 'proxyobject'. A 'proxyobject' is local resource in SALSAH which stands for an external resource. In our example, we refer to a manuscript's page of e-codices, a repository for medieval Manuscripts held in Swiss libriaries and archives (`http://www.e-codices.unifr.ch/`). Currently, there are more than a thousand manuscripts avalaible as digital facsimiles along with metadata. Recently, e-codices has also enabled a basic annotation functionality.

However, e-codices is mainly a repository offering a very attractive corpus of digitized medieval manuscripts. In order to make use of SALSAH's annotation and linkage functionality, both the search for and referencing of the e-codices manuscripts have to be possible within SALSAH. Given this possibility, the 'Narrenschiff' can be linked to a region of a page of a manuscript in e-codices (as shown in Figure 2).

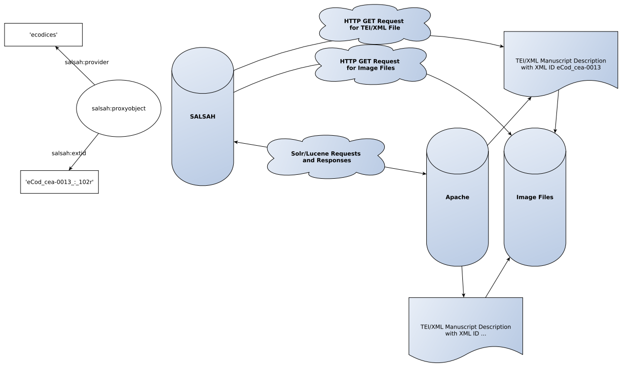

'ecodices'
salsah:provider
salsah:proxyobject
salsah:extid
'eCod_cea-0013_:_102r'
SALSAH
HTTP GET Request for TEI/XML File
HTTP GET Request for Image Files
TEI/XML Manuscript Description with XML ID eCod_cea-0013
Solr/Lucene Requests and Responses
Apache
Image Files
TEI/XML Manuscript Description with XML ID ...

Figure 5. *Relating Sources held in e-codices to SALSAH*

Figure 5 shows how e-codices is interrelated to SALSAH. A search in the SALSAH interface will not only retrieve resources locally held in SALSAH, but is also transmitted to e-codices (via a HTTP request to e-codices's Solr/Lucene search engine). The answer from e-codices is embedded in in the local search results. Once the user wants to see more details of an e-codices-resource, another request is sent to the e-codices web server providing the XML ID that identifies a manuscript (the XML ID is provided by Solr/Lucene for each record that matched the search criteria). The response to this request delivers the full manuscript description as a TEI/XML-file and the path to the digital facsimiles. When the user creates an annotation in SALSAH affecting a e-codices-resource, a 'proxyobject' is created with all information necessary to identify the resource in e-codices.

In our example, the manuscript is identified by the xml-id 'eCod_cea-0013' and the page by its pagination '102r' (resulting in `http://www.e-codices.unifr.ch/en/cea/0013/102r`). Since neither the remote images nor the metadata are stored in SALSAH, they have to be accessed when the user requests the corresponding 'proxyobject'-resource in SALSAH. This way, SALSAH makes it possible to create annotations and linkage affecting remote resources, and the generation of research data, independently of the functional range[9] of the remote host.

Summary and Conclusion

In this paper we presented SALSAH, a system for annotation and linkage of digital sources in the arts and humanities. It is not only a virtual research environment but also an infrastructure for the curation of research data. We showed how we could successfully import different collections of research data into SALSAH and make them available to researchers.

In the 'Walser' project we migrated a relational database to SALSAH. It was not a simple direct mapping of the relational structure into an RDF model. The migration process took particularly into account the researchers' domain-specific needs.

In the second project we showed how SALSAH can act as an interface between different collections of research data. Digitized medieval manuscripts stored on an external server were linked within SALSAH without transferring neither the images nor the corresponding metadata onto our server. Nevertheless, researchers can use SALSAH's full functionality even on these remote resources.

With research going on we expect SALSAH to become a useful infrastructure for the curation of research data in the humanities in the near future.

References

[1] `http://blog.pediapress.com/2014/02/visualizing-work-of-20-million.html` [accessed 2014-02-26]

[2] `http://www.library.ethz.ch/en/ms/Digital-curation-at-ETH-Zurich` [accessed 2014-03-06]

[3] Jim Gray on eScience: a Transformed Scientific Method, in: "The Fourth Paradigm, Data-Intensive Scientific Discovery", Edited by Tony Hey, Stewart Tansley and Kristin Tolle, Microsoft Research 2009, `http://www.fourthparadigm.org`, pp. XVII-XXXI. [accessed 2014-03-12]

[4] Sam Adams and Peter Murray-Rust, Chempound - a Web 2.0-inspired repository for physical science data. Journal of Digital Information, North America, 13, March 2012, `http://journals.tdl.org/jodi/index.php/jodi/article/view/5873` [accessed 2014-03-12]

[5] Lisa Johnston, A Workflow Model for Curating Research Data in the University of Minnesota Libraries: Report from the 2013 Data Curation Pilot. University of Minnesota Digital Conservancy, 2014. `http://hdl.handle.net/11299/162338` [accessed 2014-03-12]

[9] By functional range, we mean the functionality needed to create research data. Of course, the manuscript descriptions and the image files have to remain accessible on the remote host.

[6] Julia Flanders, Trevor Muñoz: “An Introduction to Humanities Data Curation”, DH Curation Guide: a community resource guide to data curation in the digital humanities, http://guide.dhcuration.org/intro/ [last updated on 2012-03-15, accessed 2014-02-26]

[7] Trevor Muñoz, Allen Renear. “Issues in Humanities Data Curation.” (2011-06-15), http://ideals.illinois.edu, http://hdl.handle.net/2142/30852 [accessed 2014-02-27]

[8] Reference Model for an Open Archival Information System (OAIS), Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book), Issue 1.1, August 2009, http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

[9] Official site of the W3C for the Resource Description Framework (RDF), http://www.w3.org/RDF/

[10] Frank Manola, Eric Miller (ed.): RDF Primer (W3C Recommendation 10 February 2004), http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

[11] Wolfram Groddeck, Barbara von Reibnitz (ed.): “Kritische Robert Walser-Ausgabe. Kritische Edition sämtlicher Drucke und Manuskripte”, Basel 2008ff.

Author Biography

Tobias Schweizer studied history, German literature and informatics at the University of Basel. Since 2010 he is an assistant and PhD student at the Digital Humanities Lab working on the development of SALSAH and on the subject of digital editing.

Andreas Wassmer received his M.Sc. in physics from the University of Zurich. He has been working as a software engineer mainly in the field of image processing and long-term archiving. His research interests are Computational Photography, High Performance Computing and the Internet of Things.

Ivan Subotic holds a Master‘s degree in Business and Economics and a PhD in Computer Science. He developed and implemented DISTARNET (DISTtributed ARchival NETwork) to redundantly store digital data in an intelligent self-reproducing distributed system.

Lukas Rosenthaler studied physics and astronomy in Basel and received his PhD also there. He worked as a Postdoc at ETH Zurich. He wrote his habilitation in the humanities department of the university of Basel about long-term archiving of digital data. Since 2012, he’s the head of the Digital Humanities Lab.