Sustainability of Linguistic Data and Analysis in the Context of a Collaborative eScience Environment

Erhard Hinrichs, Verena Henrich, Thomas Zastrow
University of Tübingen, Department of Linguistics
Wilhelmstr. 19, 72074 Tübingen, Germany
E-mail: {firstname.lastname}@uni-tuebingen.de

Abstract

For researchers, it is especially important that primary research data are preserved and made available on a long-term basis and to a wide variety of researchers. In order to ensure long-term availability of the archived data, it is imperative that the data to be stored is conformant with standardized data formats and best practices followed by the relevant research communities. Storing, managing, and accessing such standard-conformant data requires a repository-based infrastructure. Two projects at the University of Tübingen are realizing a collaborative eScience research environment with the help of eSciDoc for the university that supports long-term preservation of all kinds of data as well as a fine-grained and contextualized data management: the INF project and the BW-eSci(T) project.

The task of the infrastructure (INF) project within the collaborative research centre „Emergence of Meaning“ (SFB 833) is to guarantee the long-term availability of the SFBs data. BW-eSci(T) is a joint project of the University of Tübingen and the Fachinformationszentrums (FIZ) Karlsruhe. The goal of this project is to develop a prototypical eScience research environment for the University of Tübingen.

1. Introduction

A recent editorial in Nature magazine (Nature 461, 14; September 10, 2009) has correctly pointed out that "research cannot flourish if data are not preserved and made accessible. […] More and more often these days, a research project's success is measured not just by the publications it produces, but also by the data it makes available to the wider community.“ This observation is increasingly shared by researchers of all scientific disciplines and by funding agencies alike. At the international level, the Organisation for Economic Co-operation and Development[1] (OECD) has issued a Declaration on Access to Research Data from Public Funding, adopted on 30 January 2004 in Paris. This declaration states, inter alia, that "recognizing that an optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation" (Organisation for Economic Co-operation and Development OECD, 2004). The European Union has defined a European Roadmap for Research Infrastructures (ESFRI, 2008), which has initiated major efforts for the long-term development of sustainable research infrastructures for all areas of science, including the Humanities.

At the national level, the Deutsche Forschungsgemeinschaft[2] (DFG) has issued its Agenda 2030, which aims at making all research data that were collected with the support of public funds freely available for academic uses. As one step toward reaching this goal, the DFG encourages all collaborative research centres (German term: Sonderforschungsbereich) to create infrastructure projects (so-called INF projects) that collect all primary research data and accompanying analysis data produced within the collaborative research centre.

More generally, universities have to deal with the same issues. They need to rely on central libraries and computing service centers to address questions of long-term preservation of (primary) digital data and sustainability.

2. The Task

In order to ensure long-term availability of the archived data, it is imperative that the data to be stored conform to standardized data formats and to best practices followed by the relevant research communities. This concerns the object data as well as the metadata, which is „data about the data“ such as the creator of the resource, the time and place where the data was collected, the technical equipment, its parameters used to collect the data, etc. The International Standards Organization[3] (ISO), the World-Wide Web Consortium[4] (W3C), and the Text Encoding Initiative[5] (TEI) have formulated relevant standards for object data in the area of Linguistics. Metadata standards include Dublin Core, the TEI Header, and the best practices followed by metadata repositories such as IMDI[6] and OLAC[7].

[1] See http://www.oecd.org/
[2] See http://www.dfg.de/
[3] See http://www.iso.org/
[4] See http://www.w3.org/
[5] See http://www.tei-c.org/
[6] For ISLE Meta Data Initiative; See http://www.mpi.nl/IMDI/
[7] For: Open Language Archives Community; See http://www.language-archives.org/

Storing, managing, and accessing such standard-conformant data requires a repository-based infrastructure such as Fedora Commons[8] or the eSciDoc[9] environment that is built on top of the Fedora Commons repository. The highly flexible data management functionality of eSciDoc shown in Figure 1 is particularly suitable for a collaborative research centre such as the SFB 833. The BW-eSci(T) project will evaluate how the eSciDoc environment can be utilized as a generic solution for the entire university.

Two projects at the University of Tübingen are currently developing a collaborative eScience research environment with the help of eSciDoc: the INF project and the BW-eSci(T) project. Both these projects are introduced in the following two sections. In section 5 we then present the structure and main aspects of the eSciDoc eResearch environment. To allow the customized usage, support search functionalities and other possibilities to access data, we make use of so-called eSciDoc *solutions*. Section 6 presents customized eSciDoc solutions whose functionality is tailored to the intended user communities of the INF and BW-eSci(T) projects. Finally, a conclusion with further work follows in section 7.

3. The INF Project

There is an infrastructure project within the collaborative research centre „Emergence of Meaning“ (SFB 833). The task of this INF project[10] is to guarantee the long-term availability of the primary data, the analysis data, and the analytic tools produced by the SFB 833. What makes this task particularly challenging is the fact that the SFB will create a highly heterogeneous and possibly open-ended class of different data types and tools. The data types will include multiply annotated corpora for spoken and written language (including multi-modal data), experimental data of various kinds, including reaction time and eye-tracking experiments, self-paced reading studies, as well as electroencephalographic (EEG) and functional Magnetic Resonance Imaging (fMRI) data.

4. The BW-eSci(T) Project

BW-eSci(T)[11] is a joint project of the University of Tübingen and the Fachinformationszentrums (FIZ) Karlsruhe. At the University of Tübingen, the Department of General and Computational Linguistics (SfS) as well as the Center for Information, Communication and Media (IKM) are involved.

In the BW-eSci(T) project, the computing service center (ZDV as part of the IKM) provides the basic hard- and software as well as the system administration. The university library (also part of the IKM) integrates the eSciDoc data into its catalogues and supplies the eSciDoc installation with PIDs (Persistent Identifiers)[12]. The PID functionality is already implemented as part of the eSciDoc framework.

The SfS contributes its large collection of structured and unstructured data from the area of linguistics. To make use of this data, eSciDoc solutions need to be developed by the SfS. The FIZ Karlsruhe is responsible for adding additional functionalities into eSciDoc as needed and supports the project in any kind of technical and content related questions.

For researchers, it is especially important that all data is preserved and made available on a long-term basis. Thus the goal of the project is to develop a prototypical eScience research environment for the University of Tübingen which supports long-term preservation of all kinds of research data as well as a fine-grained and contextualized data management.

For user authentification and authorization, the University of Tübingen is a member of the DFN-AAI federation[13]. This Shibboleth-based Single-Sign-On federation guarantees that members can login to various online services, always using the same user ID and password. The eSciDoc software supports a Shibboleth-based user management so that users at the University of Tübingen and other partners in the DFN-AAI can benefit from the centralized user management.

5. The eSciDoc eResearch Environment

The eSciDoc eResearch environment has been developed specifically to be used by scientific and scholarly communities to collaborate across disciplinary and geographic boundaries. It is free Open Source software, distributed under the Common Development and Distribution License (CDDL) in version 1.0. The core functionality of eSciDoc includes a Fedora repository, a suite of eSciDoc services, and application-specific eSciDoc solutions. Figure 2 illustrates this layered architecture of eSciDoc.

The eSciDoc core (see the lowest layer in Figure 2) provides generic functionalities such as storage and versioning of data in plain file systems. The eSciDoc service layer (in the middle of Figure 2) provides basic services such as adding and updating resources, associating persistent identifiers to individual resources

[8] See http://www.fedora-commons.org/

[9] The eSciDoc project: https://www.escidoc.org/

[10] See http://www.sfb833.uni-tuebingen.de/wb/pages/de/inf-infrastrukturprojekt.php

[11] See http://www.bwescit.uni-tuebingen.de/

[12] URN (Uniform Resource Name) or handle system

[13] See https://www.aai.dfn.de/

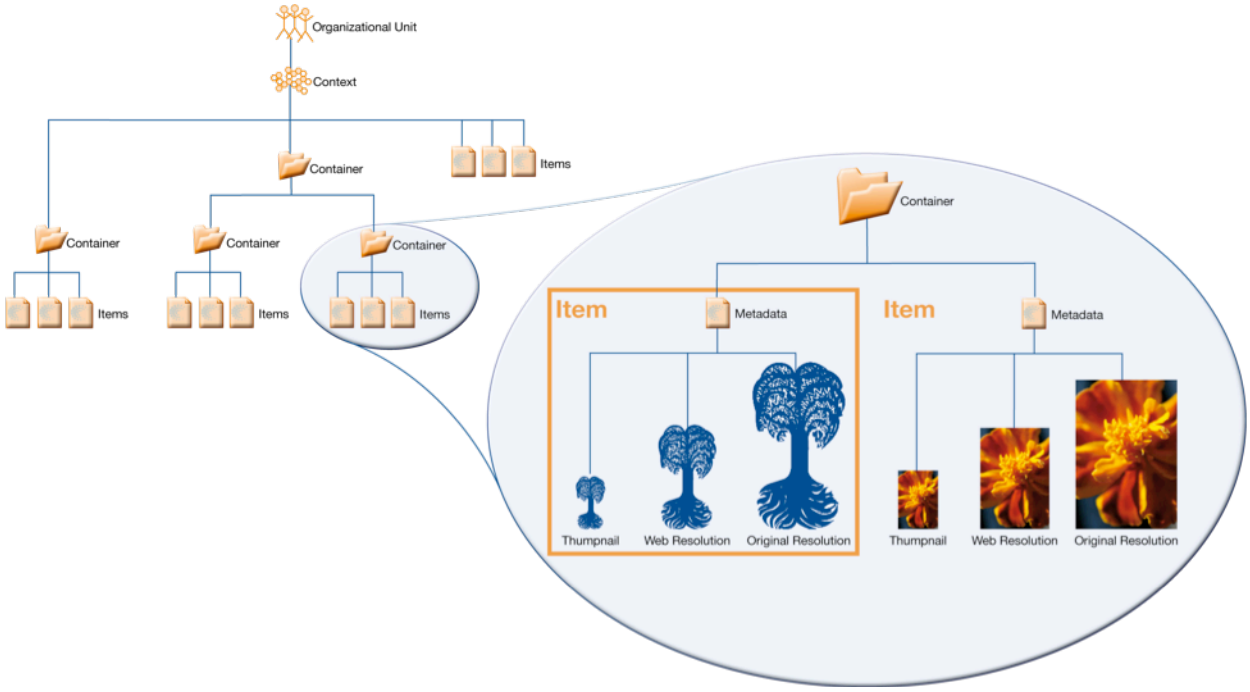

Figure 1: Fine-grained and contextualized data management in eSciDoc

or resource collections, harvesting metadata conformant with the OAI-PMH standard, and a generic search and indexing engine for object data. The top layer (shown in Figure 2), eSciDoc solutions, concerns customized applications that are particular to individual research communities.

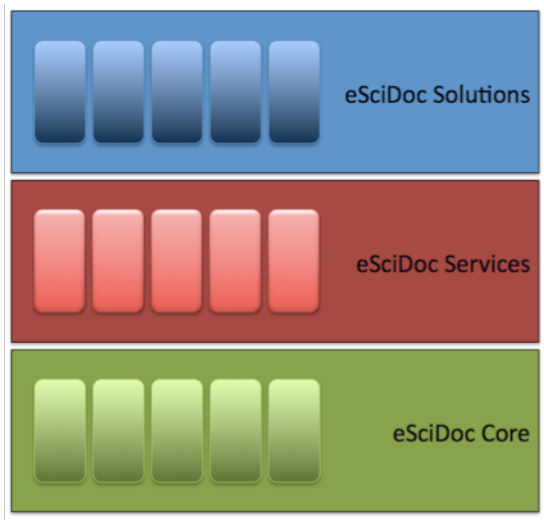

Figure 2: The layered architecture of eSciDoc

The eSciDoc system is implemented as a Java EE 5 application, which is deployed in a Java servlet container (JBoss[14]). In the background, a PostgreSQL[15] database takes care of the eSciDoc-specific metadata. The actual data, which is uploaded into the eSciDoc system, is stored as plain files in the Fedora Commons component of eSciDoc.

Figure 1 depicts the most important objects within the eSciDoc framework – *Organizational Units*, *Contexts*, *Containers*, and *Items*. Organizational Units represent the structure or hierarchy of an organization, in this case the University of Tübingen. Each faculty, department, SFB, etc. has an Organizational Unit within the Tübingen eSciDoc infrastructure. Each user is associated with at least one Organizational Unit. Context objects are associated with an Organizational Unit and represent the administrative group that owns objects within the repository and is responsible for the management of its contents and access rights. Items represent the minimal unit of content managed by eSciDoc and contain the actual data to be preserved. Items can be added directly to a Context, or can be further grouped in Containers. Thus the data itself can be stored in a recursively nested container structure. (Dreyer et al., 2007)

Both Containers and Items implement an object lifecycle, which consists of four status, each of which may have different access rights (Razum et al., 2009):

1. *Pending*/*in-revision*: In this status, only the creator can access and modify the object. This private status is assigned to an object while it is created.

2. *Submitted*: If an object is submitted, the creator cannot modify it anymore. In this status, the quality assurance takes places. The next status is either in-revision again or released.

[14] See http://www.jboss.org/
[15] See http://www.postgresql.de/

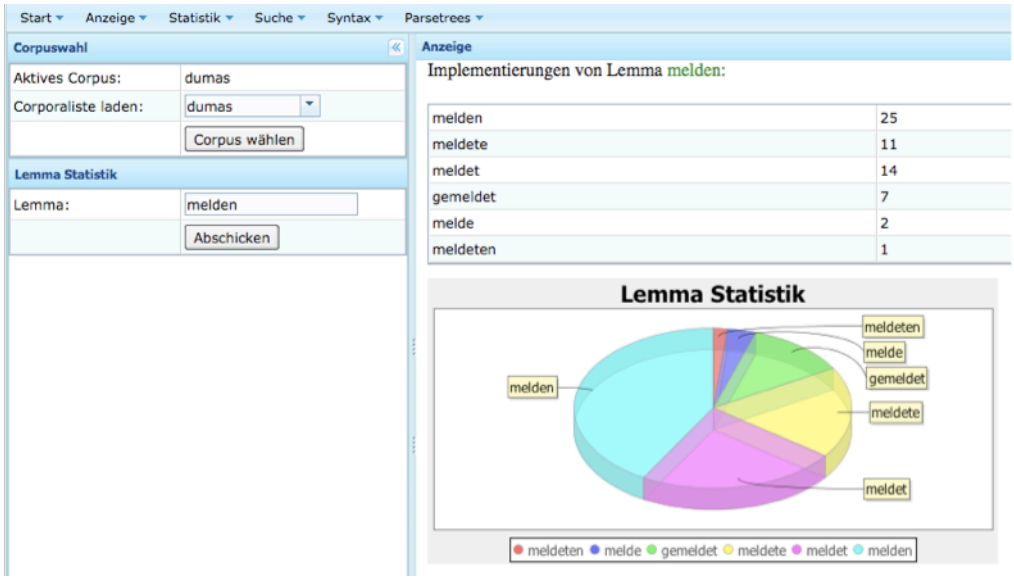

Figure 3: A customized corpus query and visualization tool

3. *Released*: Released objects are public available. They are not necessarily final but new versions can be created.

4. *Withdrawn*: This status is entered if a released object may not be publicly available anymore.

eSciDoc's service layer provides handlers for each object type. These handlers implement basic create, retrieve, update and delete (CRUD) operations on eSciDoc objects and enforce access rights. Handlers are used by solutions to interact with the eSciDoc infrastructure.

The contextualized user model, also demonstrated in Figure 1, allows privileged access rights for individual research projects within the University of Tübingen. For example, while several research projects may have access to a data set, their access rights may differ. Such access rights can also be granted to partners external to the university via the Shibboleth authentication service incorporated into eSciDoc. Access rights to eSciDoc objects can be given to individual users or to all users that belong to a particular role. A user can be assigned multiple roles, in which case he/she inherits all of the access rights associated with those roles. The use of roles simplifies the management of object access privileges.

6. eSciDoc Solutions

In eSciDoc's layered architecture, the following division of labor is intended between the eSciDoc core, generic eSciDoc services and eSciDoc solutions. Customized applications such as a corpus query and visualization tool, as shown in Figure 3, are implemented as eSciDoc solutions that make use of the generic eSciDoc search and indexing service, which in turn operates on the strongly typed data structures at the core level.

In the BW-eSci(T) project, several eSciDoc solutions are developed. These solutions concern customized applications that are particular to individual research communities, in this case for linguists. Four eSciDoc solutions for linguistic data are currently being realized:

1. The first one, which is already implemented, is a web-based search and query interface for GermaNet[16] (Kunze & Lemnitzer, 2002; Henrich & Hinrichs, 2010). GermaNet is a lexical-semantic wordnet for the German language, and this solution allows users to search for a lexeme and get semantic information about it. This includes synonyms and relations to similar sets of synonyms.

   In Figure 4, a search for the word *brauchen* (German verb for: to need) is executed and the results shown: The different readings of the word in question are all listed together with their synonyms and further details such as example sentences which illustrate the different readins.

   For a demonstration of this web-based eSciDoc solution, which uses the data of GermaNet's version 5.2 stored in eSciDoc, see http://weblicht.sfs.uni-tuebingen.de:8080/gnet/.

2. Second, a user interface for phonetic dialect data, which was acquired in the project Buldialects[17], is currently being implemented as an eSciDoc solution. It will incorporate not only the primary data of the

[16] See http://www.sfs.uni-tuebingen.de/GermaNet/
[17] See http://www.sfs.uni-tuebingen.de/dialectometry/

Figure 4: The GermaNet solution

Buldialects project, but also analysis data like matrices and topographic maps of the Bulgarian dialect distribution. The availability of electronic publications around the Bulgarian dialects will complete this eSciDoc solution.

3. The third solution differs from the first two in the type of data that it uses. While the first two solutions use eSciDoc to retrieve data, the third solution will also store dynamically created user data. This solution will be part of WebLicht[18], which stands for "Web based Linguistic Chaining Tool". With the help of WebLicht, users can create and visualize automatically annotated text corpora. This tool is accessible by a web browser, so users do not have to download or install anything on their local machines.

   Extending WebLicht to an eSciDoc solution, users will have access to personal workspaces where they can store the created text corpora online into eSciDoc. Due to legal issues and limited hard disk space, this solution will initially be available only for users at the University of Tübingen.

4. A customized corpus query and visualization tool represents the fourth eSciDoc solution. A screenshot is shown in Figure 3.

7. Conclusion and Future Work

In this paper we have presented an approach for ensuring long-term availability of the primary data, the analysis data, and the analytic tools produced by the collaborative research centre SFB 833 and at the University of Tübingen.

The goals of the INF project within the SFB 833 and the BW-eSci(T) project are very much in the same spirit as other current efforts of providing research infrastructures for humanities scholars, such as the ESFRI project CLARIN[19] and its German partner project D-SPIN. The initiative described in this paper will seek close collaboration with these projects in order to ensure compliance with the standards and policies formulated at the national and European level.

[18] WebLicht is the product of a combined effort within the D-SPIN (for: Deutsche Sprachressourcen Infrastruktur) project. For more information on D-SPIN and WebLicht, see http://www.d-spin.org/

[19] For: Common Language Research Infrastructure Network; http://www.clarin.eu/

8. Acknowledgements

Acknowledgements go to our BW-eSci(T) colleagues of the Fachinformationszentrum Karlsruhe especially for their help and support on the eSciDoc system. Further thanks go to our local partners at the Center for Information, Communication and Media of the University of Tübingen for the close collaboration on the infrastructural aspects of eSciDoc.

9. References

Dreyer, M., Bulatovic, N., Tschida, U., Razum, M. (2007). eSciDoc – a Scholarly Information and Communication Platform for the Max Planck Society. In *German e-Science Conference*.

European Strategy Forum on Research Infrastructures ESFRI (2008): *European Roadmap for Research Infrastructures*. http://cordis.europa.eu/esfri/roadmap.htm Last update on: 16 April 2008.

Henrich, V., Hinrichs, E. (2010). GernEdiT – The GermaNet Editing Tool. *Proceedings of LREC 2010*, main conference.

Kunze, C., Lemnitzer, L. (2002). GermaNet – representation, visualization, application. *Proceedings of LREC 2002*, main conference, Vol V. pp. 1485-1491.

Nature 461, 145; September 10 (2009) : *Data's shameful neglect*. Editorial, doi:10.1038/461145a http://www.nature.com/nature/journal/v461/n7261/full/461145a.html

Organisation for Economic Co-operation and Development OECD (2004): *Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique*; ANNEX 1: DECLARATION ON ACCESS TO RESEARCH DATA FROM PUBLIC FUNDING, adopted on 30 January 2004 in Paris: http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

Razum, M., Schwichtenberg, F., Wagner, S., Hoppe, M. (2009). *eSciDoc Infrastructure: A Fedora-Based e-Research Framework*. M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 227-238.