Sustainability of Linguistic Data and Analysis in the Context of a Collaborative Research Center

Erhard Hinrichs, Thomas Zastrow, and Verena Henrich
Seminar für Sprachwissenschaft, Univerisität Tübingen
{firstname.lastname}@uni-tuebingen.de

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 (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." The European union has defined a European Roadmap for Research Infrastructures, 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 (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 centers (German term: Sonderforschungsbereich) to apply for infrastructure projects (so-called INF projects) that collect all primary research data and accompanying analysis data produced within the collaborative research center.

2 The Task

There is such an infrastructure project within the collaborative research center „Emergence of Meaning“ (SFB 833). The task of this INF project 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 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.

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. 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. International Standards Organization (ISO), the World-Wide Web Consortium (W3C), and the Text Encoding Initiative (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 and OLAC. Storing, managing, and accessing such standard-conformant data requires a repository-based infrastructure such as Fedora Commons or the eSciDoc Environment that is built on top of the Fedora Commons repository.

3 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. Its core functionality includes a Fedora repository, a suite of eSciDoc services, and application-specific eSciDoc solutions. Fig. 1 illustrates this layered architecture of eSciDoc.

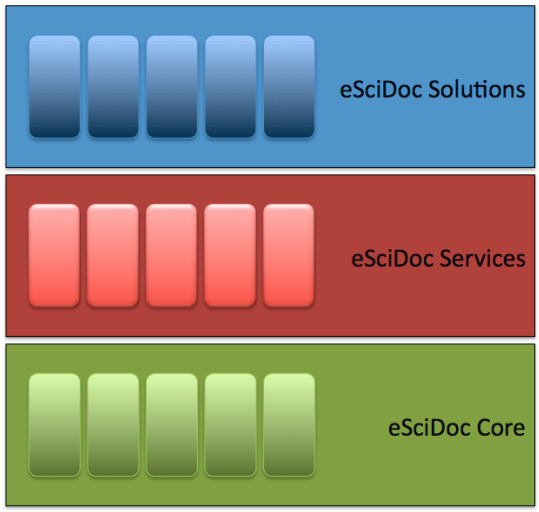

Figure 1: The layered architecture of eSciDoc

The eSciDoc core provides generic functionalities such as storage and versioning of data in plain file systems. The eSciDoc service layer provides generic services such as associating persistent identifiers to individual resources or resource collections, a metadata harvesting conformant with the OAI-PMH standard, and a generic search and indexing engine for object data. The third layer, eSciDoc solutions, concerns customized applications that are particular to individual research communities.

The highly flexible data management functionality of eSciDoc shown in Figure 2 is particularly suitable for a collaborative research center such as the SFB 833.

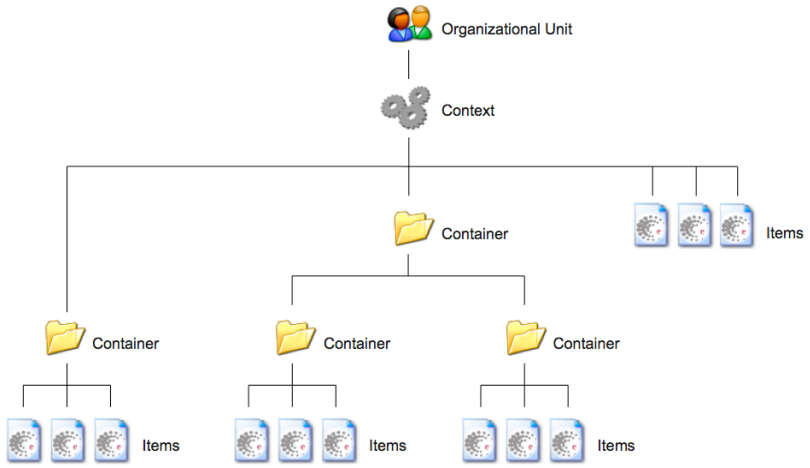

Figure 2: Fine-grained and contextualized data management in eSciDoc

The contextualized user model allows privileged access rights for individual research projects within the SFB 833. For example, while several research projects may have access to a data set that was produced by another research project, their access rights may differ. Such access rights can also be granted to partners external to the SFB 833 via the Shibboleth authentification service incorporated into eSciDoc. The data itself is stored in a recursively nested container structure.

When we apply eSciDoc's layered service architecture to the SFB 833, the following division of labor is put in place 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 full paper, a wider range of customized eSciDoc solutions such as query tools for multi-layered linguistic annotations and rendering tools for neuroimaging data will be discussed.

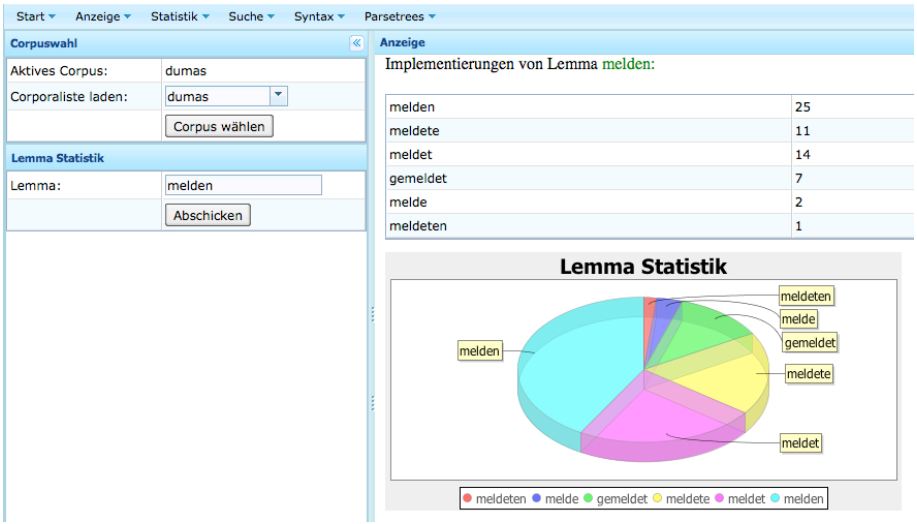

Figure 3: A customized corpus query and visualization tool

4 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 center SFB 833. The goals of the INF project within the SFB are very much in the same spirit as other current efforts of providing research infrastructures for humanities scholar, such as the ESFRI project CLARIN (for: Common Language Research Infrastructure Network; ww.clarin.eu) and its German partner project D-SPIN (for: Deutsche Sprachressourcen Infrastruktur; www.d-spin.org). The SFB 833 initiative described here will seek close collaboration with these projects in the near future in order to ensure compliance with the standards and policies formulated at the national and European level.

References

European Roadmap for Research Infrastructures, issued by the European Strategy Forum on Research Infrastructures (ESFRI): http://cordis.europa.eu/esfri/roadmap.htm

The eSciDoc project: https://www.escidoc.org/

Nature 461, 145; September 10, 2009; doi:10.1038/461145a http://www.nature.com/nature/journal/v461/n7261/full/461145a.html

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