LAUDATIO-Repository.org – A Modular Approach of an Open Access Research Data Repository for a discipline

Dennis Zielke

Computer and Media Service
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
zielkede@cms.hu-berlin.de

Abstract: The presentation aims at presenting the LAUDATIO repository, a data repository for research data of historical corpus linguistics. It has a technical focus concerning the requirements that have been taken into consideration for the development of a repository infrastructure based on Fedora.

Research data collections are expensive in terms of time, money and knowledge that is necessary to compile them. The scientists' effort can be reduced by sharing these collections with a broader scientific community for re-use and re-evaluation. A common approach of creating sustainable and shareable data is the definition of existing and future open standards for the data creation and maintenance. The implementation and usage of digital repositories enables the management and distribution of that data. However, large collections of linguistic data, like corpora of deeply annotated text, pose unique challenges to repository design, which mainly arise due to the complexity of the applied data models. In the presentation a historical corpus linguistic data repository that is based on community-specific requirements and distinct user scenarios will be presented. Its main focus is on German historical texts including all dialects of time periods ranging from the 9th to the 19th century.

The LAUDATIO repository is developed in the cause of a research data infrastructure project funded by the German Research Foundation from 2011 until 2014 at Humboldt-Universität zu Berlin within a cooperation of Computer and Media Service, the Departments of Historical Linguistics and Corpus Linguistics at HU Berlin, as well as INRIA, France. [1]

The main objective for the technical implementation of the repository was to ensure the long-term access and preservation as well as the accessibility and the (re-)use of the heterogeneous research data. The requirements specification is essentially based on criteria such as the use of a complex data model for linguistic research data [Od13], their management, the structural collocation, the searchability, and long-term availability of the data. [Zi14]

Exemplary technical requirements regarding these aspects and the implemented features of the repository infrastructure include:

[1] LAUDATIO Repository: http://www.laudatio-repository.org [last access 26.06.2014]

- Compatibility with a complex data model
- Data indexing and search
- Data Versioning
- Registration and management of Persistent Identifiers (PID)
- Graphical User Interface (GUI) for data presentation
- Search and visualization with the help of ANNIS[2]

Fedora Commons[3] being an open-source repository framework that is well established in the repository community turned out to be the well-suited solution for the storage and management of linguistic research data considering the defined requirements. In the digital humanities domain Fedora is a widespread repository solution, and is for instance recommended and used by the CLARIN research infrastructure and its participating institutions. The LAUDATIO repository infrastructure is furthermore based on generalizable software modules such as the graphical user interface, the data exchange module between research data and Fedora REST-interface[4], and an indexed and faceted search based on Lucene-based[5] technology.[6] Naturally, the repository is oriented towards international guidelines concerning organizational, legal as well as technical aspects of research data repositories such as the Data Seal of Approval[7].

The presentation aims at a discussion on the re-usability of the LAUDATIO repository infrastructure modules for other initiatives providing research data infrastructures as well.

References

[Od13] Odebrecht, C.: Modellierung von linguistischen Forschungsdaten. Korpuslinguistik Kolloquium 13.11.2013, Berlin., https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen-en/carolin/modellierung-von-linguistischen-forschungsdaten/view

[Zi14] Zielke, D.: Open Access Research Data Repository for Corpus Linguistic Data – A Modular Approach Approach.Open Repositories 2014 Conference. 9.-13.6.2014 Helsinki, Finland, http://www.laudatio-repository.org/laudatio/wp-admin/tmp/2014/04/LAUDATIO-OR2014.pdf.

[2] ANNIS: http://www.sfb632.uni-potsdam.de/annis/ [last access 26.06.2014]
[3] Fedora Commons: http://www.fedora-commons.org/ [last access 26.06.2014]
[4] Representational State Transfer
[5] Lucene: An open source full-featured text search engine library, http://lucene.apache.org/core/ [last access 26.06.2014]
[6] Technical documentation of the LAUDATIO repository: http://www.laudatio-repository.org/repository/technical-documentation/; Sourcecode: https://github.com/DZielke/laudatio [last access 26.06.2014]
[7] Data Seal of Approval: http://datasealofapproval.org/en/ [last access 26.06.2014]