Archiving with Quality - implementing a certified digital archive

We present ACDH-OeAW's new digital archiving system ARCHE concentrating on crucial steps in its making. ARCHE is one of the central services of ACDH-OeAW and aims to offer stable hosting of digital research data for the Austrian humanities community.

While ARCHE's predecessor, CLARIN Centre Vienna / Language Resources Portal (LRP) (https://clarin.oeaw.ac.at), was focused on digital language resources, ARCHE is designed to cater to the needs of a more diverse audience accepting a broader range of data created as part of research in the humanities and cultural/social studies. ARCHE is intended to become a certified trustworthy archiving system complying with the requirements of the Data Seal of Approval. This aspiration demands not only a solid technical basis for storing and preserving, but also a carefully crafted metadata schema allowing for efficient discovery and dissemination of the resources, an easy to use interface to search for and reuse data and a set of policies and documentation, making transparent the underlying design decisions and available functionalities.

The heterogeneity of data represents a major challenge for the development of the repository, especially in terms of supported metadata schema(s). Metadata has to be generic and also specific enough to describe the variety of resources. Another important aspect in modelling the metadata schema is reuse of and/or compatibility with established metadata schemas, which in some cases are very discipline specific.

An initially compiled schema that reused well known metadata schemas like DCMI Metadata Terms, FOAF (friend of a friend), SKOS (Simple Knowledge Organization System) or RDFS (Resource Description Framework Schema), in combination with own newly introduced properties, lead to a construct which was difficult to handle both for authoring and curation of metadata. Consequently, we adopted an alternative approach, in which we developed our own schema based on metadata used in a representative selection of resources to be deposited in the repository. To ensure delivering information in established metadata formats, equivalencies between the properties of our schema and properties from established schemas were defined.

Regarding the design of the technological stack for ARCHE, we have evaluated a number of existing solutions. In the end, we have opted for Fedora Commons, a widely used system, also employed by other Austrian service providers (GAMS and PHAIDRA), and for implementing LRP. However all current solutions still rely on version 3 of Fedora Commons, which reached end of life in 2015. Consequently, when designing the new repository, we decided to adopt Fedora version 4, which entailed substantial development effort, due to backward compatibility issues. ARCHE's software architecture comprises an array of software components. The repository itself consists of numerous components, especially also a triple store (Blazegraph) as the persistence layer for the metadata. On top of that we developed custom components in PHP: Ingestion is accomplished semi-automatically with ingestion scripts relying on a utilities library. The same library is also used as the basis for a flexibly configurable OAI-PMH endpoint for serving metadata, and for a user interface allowing to search and browse through the resources' metadata, as well as for inspecting the resource descriptions and accessing the actual resources. This generic interface is accompanied by a growing set of dissemination services that are able to represent specific data types in various forms and formats, examples being HTML renditions of TEI documents, rendering of geodata on maps or graph data as visual interactive networks.

To ensure quality and relevance of contents, deposition of data will always take place in interaction with the repository curators and will adhere to the Collection Policy in place. Data will go through a standardised workflow according to the Open Archival Information System (OAIS).

At the beginning of the deposition process, data undergoes a number of automated checks to ensure its formal quality. The results of the automatic checks serve as input for manual curation process including examination and enrichment of metadata, necessary format conversions, and clarifying legal issues. At the end of the process, depositors will have to sign a Deposition Agreement.

We are planning to submit the new repository for assessment both for the Data Seal of Approval as well as for CLARIN Centre B status in autumn 2017.