Ecosystem for digital data preservation

Author: Raman Ganguly
Affiliation: University of Vienna

Keywords: data presentation; repositories; workflow; research data; data life cycle

Abstract

This presentation will address issues concerning the handling of complex data such as research data, multimedia content, e-learning content, and the use of repositories. At the University of Vienna an ecosystem for digital data preservation is set up and will be enlarged in the future. This ecosystem is the foundation for research data management and it is established as a central service. For this central services the FAIR principal (Findability, Accessibility, Interoperability and Reusability) are the main driver.

Based on a ten year experience a four phase workflow model of for digital data presentation was establish to address the roles and the responsibilities in the process. This was necessary because the process for running the university repository Phaidra shows that it is unclear what the service of data presentation provides. With this model we could now present our main field of action and hence focus on the upcoming topics like the complexity for research data management. We could also see that this model works for other institution by shared it over the Phaidra partner network.

In 2014 the project e-Infrastructures Austria was started and the University of Vienna put the project management in place. e-Infrastructures Austria was a project for the coordinated establishment and development of repository infrastructures for digital resources in research and science throughout Austria to securely archive and publish digital publications, multimedia objects and other digital data resulting from research and education. During one work package of this project a survey was undertaken within the scientific community in Austrian universities and external university research institutions. This survey helped to identify the status quo on how research data is currently handled.

For us the results directly leads to our repository infrastructure and we started to think about how we can operate the complexity of the different needs. Influenced by micro service architecture and the concept of open data we decided to establish an ecosystem of multiple different services which closely works together and can service the hole research process as well as the data life cycle.

The initial point was Phaidra our longterm archiving system for generic data where we can store many different kinds of data. In this repository all metadata and controlled vocabularies were administrated. Per definition the initial starting point was Phaidra, our longterm archive with a persistent identifier, no data can be deleted. In the data life cycle it is also necessary to delete data, special when repositories should be a part of e-learning. For this we run now a second repository for midterm archiving where data can be deleted and in the future also seamlessly pushed to the longterm archive. For managing the controlled vocabularies a terminology server is implemented based on the SKOS standard.

The survey also shows that we have to deal with special kinds of data like data bases and software. For some of them good solution are already well establish like git for software management. We integrated a git services based on GitHub to our ecosystem and we are able to link the software releases to their data in our repositories. A persistent identifier service, which we realized with Handle System, can now also use a persistent identifier for a revision in the GitHub repository.
Handle is open source and it is the underlying technology for DOI (Digital Object Identifier).

In the future we plan to integrate a service for data management plans based on the DMPOnline Tool from DCC (Digital Curation Centre based at the University of Edinburgh) and the recommendation from RDA (Research Data Alliance) regarding active data management plans. These are data management plans which are provided in human and also in a machine readable way. The machine readable output can further be used in tools for data stewardship. This gives the possibility to have more control of the data, there provenance and context. Those are relevant information for the reuse.

Next steps on the roadmap are services for data bases. In this case we first have to define what we can cover with standard services. Requirements engineering in this case has already been started to see what features will meet the need of researchers. So far we see that there are big differences between research disciplines in the meaning of the terminology “data base”. E.g. historians often include the representation layer when discussing data bases.

Currently a manifold of software development takes place which is a big challenge regarding infrastructures as well as coordination. The question is if software developed by research projects is a part of data preservation and if yes how can this be maintained after the project ends. New concepts have to be implemented and we also see a big need in technical consulting for managing the data.

Important steps are for this project are not only to provide a good working infrastructure, it is essential to get connected with the research community and have a link to other infrastructure projects. That is the reason why our service is linked to projects like OpenAIRE and Europeana and we are in regular contact with GÉANT. We also observe the European Open Science Cloud and the big Austrian infrastructure projects like the Vienna Scientific Cluster.

**Speakers biography**

Raman Ganguly is the leader of the Software Design and Development Department at the Vienna University Computer Center (ZID), and is an expert in software development and media technology. Since 2011, he has devoted his energies to the digital archivation of data from a technical perspective with a focus on architectonic models for technical implementation. His duties at the Vienna University include the technical development and operation of repositories, as well as the university website. For him, modular systems and consumer-oriented representation of data are the most important aspects of constructing modern IT landscapes.

Raman Ganguly is the technical leader of Phaidra, member of the Phaidra Steering Committee and coordinates the technical agenda of the international consortium www.phaidra.org. Since 2014, he has been the technical leader of the federal-grant (HSRM) project e-infrastructures Austria and since 2016 also the technical leader of the federal-grant project Open Education Austria.