hvw/2016-08-08

Proposal for a talk (~15 min) at CKANCon 2016, October 4th, Madrid

CKAN for Research Data Management at Eawag:
A report from the trenches

Harald von Waldow <harald.vonwaldow@eawag.ch>
Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

Abstract

Eawag employs about 500 staff in research, teaching and consulting in twelve research departments that span natural sciences, engineering, and social sciences. Eawag is concerned with concepts and technologies for dealing sustainably with water bodies and with water as a resource. Integral part of its mission is to provide a link between basic science and practical applications.

Accordingly, the research data produced is very diverse in every aspect. We are dealing with dataset volumes from kilobytes to hundreds of gigabytes. Anticipated time-horizons for storage range from a couple of years to practically infinite "long-term" archival. The original data may take any form, e.g. CSV tables, video files, baroque Excel spreadsheets, or an SQL database dump. The content ranges from locally stored questionnaires with sensitive personal data to large transcriptomic datasets that are (also) stored in field specific off-site databases.

Eawag has decided to establish an institution-wide central research data repository to achieve several aims:

- Enable compliance with the Guidelines for Good Scientific Practice (e.g. reproducibility).
- Archival of unique unrepeatable measurements ($\approx$ everything in the environment).
- Facilitate data management / data publication requirements by funding agencies and publishers.
- Foster in-house data discovery and collaboration across working groups and departments.
- Facilitate intra-workgroup data management workflows.

We chose CKAN for a reason as a basis for this repository, but technical challenges remain. Some are rooted in the discrepancy between the main area of CKAN, open data publishing, and our requirements. I will present some of the main challenges we are facing, the progress we made so far through customizing and extending CKAN, ideas and strategies to deal with remaining problems, and an outlook to the feature set we envision to have available when the system goes into production.

Some of the characteristics and requirements of the project are:

- Focus on easy data submission and quality control: The submission procedure is supposed to be quick, simple and easily communicated to serve a transient user base (think PhD students) with limited tolerance for additional workload (staff scientists). In this respect, technical measures can only go so far and support, training and communication is of the essence.
- Tools for automated data preparation, annotation and upload.
- A meta data scheme (or several) that makes sense for all users and is consistent with established standards for research data to allow for interoperability with other repositories.
- A search interface that (also) allows for field specific queries from experts.
- Versioning and integrity checks for resources and datasets.
- Access restriction management for sensitive data.
- Integration with the institutional publication repository.
- Facilitation of the registration of Digital Object Identifiers (DOIs).