# Towards a distributed research data management system

M. Politze[1], F. Krämer[2]

[1]IT Center RWTH Aachen University, Seffenter Weg 23, 52074 Aachen, politze@itc.rwth-aachen.de
[2]IT Center RWTH Aachen University, Seffenter Weg 23, 52074 Aachen, kraemer@itc.rwth-aachen.de

**Keywords**
RDM, distributed systems, PID, metadata, RDF

## 1. SUMMARY

At RWTH Aachen University a project aims at improving the support and technical infrastructure for Research Data Management (RDM). In this project the need was identified to provide researchers with a tool to simply register and store metadata corresponding to their files. Our solution enables researchers via a web interface to register and identify their data with PIDs as well as store metadata in a standardized form. To account for confidentiality concerns data and metadata can optionally also be stored in a file locally. The solution will be deployed and evaluated in March 2016.

## 2. BACKGROUND

There is an initiative to set up an integrated Research Data Management (RDM) system within the next years at RWTH Aachen University. A project group focuses on consulting and training as well as on the development of technical solutions for RDM. Since managing data requires extra effort from researchers, usability and seamless integration into existing workflows are key to establishing an integrated RDM. Technical solutions need to cover all domains of the research process: private and collaborative domain, in which researchers actively work with the data, as well as the archive and publication domain, in which data is accessed less frequently. Registering metadata is the prerequisite for re-discovering and re-using data, but requires extra effort from researchers. However, this task is the easier the earlier it is performed since the context of the data generation is more present to the researcher. We therefore want to offer a tool for the registration of metadata that works independently from the local IT environment of the researcher, allowing every researcher to document metadata as soon as research data is produced. It also needs to be flexible regarding the storage of data as well as metadata since many institutes have their own IT infrastructure and sometimes strong privacy and confidentiality concerns regarding even meta-information about their research. Furthermore, structure and format of metadata need to be standardized to allow for a seamless transition to the archive or publication domain and for the data to be found and re-used later. The RDM project closely cooperates with a number of pilot users, which will test the solution throughout the development process.

## 3. OUR SOLUTION

Our first step towards an integrated RDM system is to offer a distributed solution that allows the usage of an integrated service layer (ISL) as well as any private services already available to the researcher.

Researchers log on to a web interface where they can register their data to a PID (persistent identifier) service. They can choose from a number of predefined metadata schemas that use Dublin Core as a basis, and fill it in. Depending on the user's institutional context it is also possible to preset a metadata schema. To achieve maximum re-usability the metadata schemas and presets are defined in RDF format. The link between data and metadata will be established using the registered PID. Via a REST API it will also be possible to automatically provide metadata from local systems such as test stations or measurement devices. Storage of the data and metadata can be realized in two different workflows depending on the researcher's preferences.

(1) Private Workflow: Formalized metadata is stored together with the research data in the private archive. Metadata is provided in an RDF file. To ensure retrievability data and metadata are linked by a PID service. (2) Integrated workflow: Formalized metadata is stored in the integrated metadata store. With the number of users and disciplines the available metadata schemas will be continually expanded. The storage concept for metadata still poses a challenge. For the first version a generic RDF triple store will be used. The researcher also has the option to transfer the research data into the integrated archive service. A PID links data and metadata, regardless of where the data is kept. Already existing metadata can be directly imported into the metadata store as long as it is formalized and the metadata schema is known. This process can be further automated using the REST interfaces to update and copy metadata in the ISL. The interface also allows updating links stored in the PID service in case of server migrations (private to private, private to integrated, integrated to integrated).

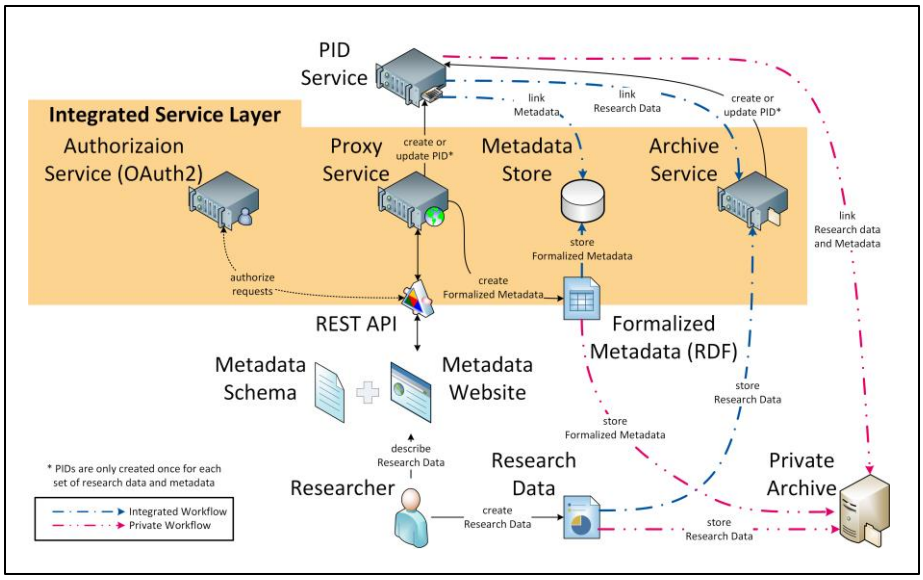


**Figure 1:** Architectural overview of the first stage

The Web Page uses W3C recommendations for defining metadata schemas and generates a user interface to enter the metadata. The formalized metadata is stored in RDF format to ensure long-term usability.

## 4. CONCLUSION AND OUTLOOK

The shown architecture allows multiple integrated and private archives which help supporting (1) different archive types in the integrated layer e.g. short- vs long-term or high vs low latency and (2) the gradual adoption by researchers of the ISL systems instead of their own infrastructure. Also this allows the ISL systems to slowly grow and meet the expected functionalities. Automation features are a key feature for adoption by power users with large data volumes.

In the world of IT long-term storage of ten or more years is quite challenging. Storage of formalized metadata in RDF format ensures long-term usability and facilitates migration between successive systems and retrievability using advanced search algorithms. However, this requires the specification and usage of existing metadata standards.

As a minimum set of metadata DublinCore is used. Finding the right discipline specific metadata schemas has turned out to be difficult. Discipline specific metadata schemas have yet to be analyzed and evaluated together with data curators from the university library and researchers from the discipline.

The current solution will be deployed in March 2016 and evaluated by users from different disciplines. According to the user feedback it will be adapted and further extended to meet all requirements currently posed by the researchers.

## 5. AUTHORS' BIOGRAPHIES

**Marius Politze, M.Sc.** is research assistant at the IT Center RWTH Aachen University since 2012. His research is focused on service oriented architectures supporting university processes. He received his M.Sc. cum laude in Artificial Intelligence from Maastricht University in 2012. In 2011 he finished his B.Sc. studies in Scientific Programming at FH Aachen University of Applied Sciences. From 2008 until 2011 he worked at IT Center as a software developer and later as a teacher for scripting and programming languages.

**Florian Krämer** studied Political Science, Economics and Linguistics and received his Master of Arts from RWTH Aachen University in 2010. After working as a research assistant in the Institute for Political Science he joined the IT Center in 2011. Here his tasks first included support and training, he was responsible for the online documentation and worked on different projects including knowledge management and research data management. Since 2015 he is responsible for the coordination of the activities concerning RDM within the IT Center and a member of the RWTH project group on RDM.