simpleArchive – Making an Archive Accessible to the User

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
eScience, RDM, long term archive, service oriented architecture, PID

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 upload and save files in a long term archive. Our solution allows the researchers to use a web interface to deliver their data as a single file into a tape archive. All uploaded artifacts are identified using a PID.

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 (Eifert, Muckel, & Schmitz, 2016). 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.

Long term data storage is becoming more and more important as research funding organizations require researchers to make their data available and re-usable (RfII – German Council for Scientific Information Infrastructures, 2016). At RWTH Aachen University archiving data using the IBM TSM client requires technical expertise that not every researcher has. Therefore, the goal of simpleArchive is to allow researchers to access long term archival capabilities without technical knowledge by way of a Software as a Service (SaaS) that uses state of the art web technologies. Archiving a file for the user is as simple as uploading a file on a web page.

3. OUR SOLUTION

The web interface presented by simpleArchive can be used to upload the research data to be archived. The data is then temporarily stored on a server before being transferred into a tape archive where the data can be stored for long terms at relatively low costs. The actual archival process is transparent to the user. Immediately after the user has completed the upload of the data a PID is issued which in turn may be used for referencing the data in a text publication or on a web page.

To restore data from the archive the researchers use the PIDs issued to identify their data. Using simpleArchive they can request to restore the data which is then copied from the tape archive to a temporary file store. Based on the temporary file a download URL is generated that the researcher can use to access the archived data. The URL is provided using the download mechanisms of GigaMove (Bischof, Bunsen, & Hinzelmann). This especially means that the restored data is only accessible temporarily using the URL. Afterwards the restoring process needs to be triggered again.

Additionally to the uploaded data, the researcher's affiliated institution is saved. If the researcher retires or leaves the university, it is then possible to name the institution responsible for the data. This is especially important since long term archives likely need to be migrated several times. Therefore it is desirable to save only data that is still relevant. A full process defining how data can actually be evicted from the archive, however, remains to be discussed.

Furthermore, if the PID is obtained by other researchers an anonymous landing page allows them to get in touch with the researcher who archived the data. Using the restore process described above the download URL can then be passed to allow access to the data within a community of researchers.

The web services used by simpleArchive are integrated into the existing infrastructure for personalized web services (Politze, Schaffert, & Decker, 2016). It is furthermore becoming part of an integrated service layer supporting the researcher throughout the research process. The integrated service layer depicted in Figure 1 merges different IT services used by the researchers.

4. CONCLUSION AND OUTLOOK

Basing the API on the processes rather than the underlying services effectively reduces the impact of vendor lock in. This design allows to add new services or to scatter requests between multiple services based on certain rule sets. This actively decouples the systems into smaller functional units, which in turn increases maintainability. Furthermore, this allows granting access to the service as part of a cooperation with other institutions or universities.

A generic metadata web interface is also part of the integrated service layer and allows the researcher to describe the uploaded data. Currently the collected metadata is saved as a RDF file. While this file can be downloaded and reviewed by the researcher the whole metadata library is not queryable as such (Politze & Krämer, 2016). A future project will consider different techniques, using triple stores to create an interactively queriable interface for the researchers.

This service oriented architecture organizes the access to the backend systems using vendor specific or legacy APIs. Basing the API on the processes rather than the underlying information services effectively reduces the impact of vendor lock in and therefore makes future migrations to proceeding systems easier. Furthermore, this design allows to add further information services or to scatter requests between multiple systems based on certain rule sets. This actively decouples the systems into smaller functional units, which in turn increases maintainability.

5. REFERENCES

Bischof, C., Bunsen, G., & Hinzelmann, S. (n.d.). Gigamove – Einfach und schnell große Dateien austauschen. In Verein zur Förderung eines Deutschen Forschungsnetzes e.V..

Eifert, T., Muckel, S., & Schmitz, D. (2016). Introducing Research Data Management as a Service Suite at RWTH Aachen University. In P. Müller, B. Neumair, H. Reiser, & G. Dreo, *9. DFN-Forum Kommunikationstechnologien* (pp. 55–66). Bonn: Gesellschaft für Informatik e.V. (GI).

Politze, M., & Krämer, F. (2016). Towards a distributed research data management system. In Y. Salmatzidis. Thessaloniki.

Politze, M., Schaffert, S., & Decker, B. (2016). A secure infrastructure for mobile blended learning applications. In J. Bergström, *European Journal of Higher Education IT 2016-1*. Umeå.

RfII – German Council for Scientific Information Infrastructures. (2016). Performance through Diversity - Recommendations regarding structures, processes, and financing for research data management in Germany. Göttingen.

6. AUTHORS' BIOGRAPHIES

Marius Politze, M.Sc. is research associate 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.