On the Decentralization of IT Infrastructures for Research Data Management

Marius Politze[1], Thomas Eifert[1]

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

Keywords
RDM, decentralized infrastructure central components, process

1. ABSTRACT

In spite of a wide variety of generic IT services offered to researchers to enhance their data management, publishing, keeping public records and providing long-term storage of research data are often unsolved problems for researchers (Dreyer & Vollmer, 2016). On the other hand, it can be seen that specialized solutions that are well adapted to the needs of researchers have a high level of acceptance(Curdt, et al., 2016). Tailoring an IT infrastructure to the specific requirements of certain disciplines thus is crucial for their short-term success. If data management is viewed as a university wide and generic task, the benefit for researchers is often indirect. The integration into research processes is a central challenge when establishing IT support for a data management system.

RWTH Aachen University and many other universities have a highly decentralized technical infrastructure. This is particularly important when data is exchanged between systems that have to be connected (Eifert & Bunsen, 2013). IT services supporting research data management originate from central institutions such as university libraries, computing centers or IT departments, but also from individual institutes or chairs, where researchers themselves operate services for their own. The result is a manifold, decentralized IT system landscape.

Central processes and offers that are so deeply integrated into the daily work processes of researchers must not only consider this decentralization, but also actively deal with it in order to be attractive for the users and offer benefit. Concerning decentralized services, sustainability is becoming more important, especially for publication and long-term storage of research data, and equally for the researchers' processes to rely on the central components. In organizational terms, this is often the responsibility of central service providers: when the work on a research project has been completed, but increasingly also for interim results to be published, there is thus a transfer point between centralized and decentralized services. In order to support this data life cycle, the systems used must enable this transfer as seamlessly as possible, to blur this transfer point.

Hence, two points of the research data life cycle are critical: the time of data archival at which all information about the data set should be available, and the time of data generation at which the information is available. The longer these points lie apart, the more complicated it is to obtain missing information. In other words, there is information available in the moment of data generation which should be conserved as soon as possible in the data life cycle and to associate it explicitly with the data set. This meta data should be recorded in a form that allows direct use for later publication.

In the field of data management, retention periods of ten years or more are quite common. For the IT infrastructure, however, this period is quite a challenge: Typical maintenance contracts for server hardware, run for five years and fast moving software life cycles, require that data repositories have to be migrated about three times between successive systems during their retention time. Hence, services supporting data management must allow that preserved data can be migrated. It is therefore inevitable to plan how systems can be replaced or even be switched off completely. Because the degree of integration of systems in the research processes, the interchangeability or, from a certain perspective, the "exit scenario", must therefore be an integral part of the initial consideration.

From our perspective, the most promising way to deal with this problem is to define technology-independent and process-oriented interfaces. Instead of specific formats and protocols, an independent and open specification is used that is oriented towards supported processes. This abstract

intermediate layer is responsible for the translation between processes of researchers and the implementations of manufacturers. This type of process-oriented modelling is in line with current standards in software architecture such as Web services and service-oriented architectures (Dumas, van der Aalst, & ter Hofstede, 2005). For IT components, this approach allows to use proprietary commercial solutions, e.g. for the storage layer, due to the interchangeability built in by our approach.

Some existing services can be used to support the decentralized scenarios outlined above with a central service portfolio. For the implementation of the decentralized data life cycle at RWTH Aachen some basic technologies are combined and integrated into individual research processes.

Persistent identifiers (PIDs) of the handle system are used as the basic technology for identifying data records. The service is provided by the European Persistent Identifier Consortium (ePIC). The PIDs allow the data record to be identified and referenced before it is published, locally within the research group or globally with collaboration partners outside the university. In addition, at least upon request, the underlying data set can be identified and retrieved in the local context of the research institution.

In addition to the description of the research data, long-term storage plays an important role at the end of the data life cycle. Data is no longer changed and the access frequency decreases. A tape archive is particularly suitable for data that is not published but should be kept in case it is to be reused. The central tape archive of the RWTH Aachen comprises over 900TB of data in December 2018, of which 200TB were delivered in 2017, which shows the increasing importance of this technology. In order to make this technology more accessible to researchers, a simplified workflow was established with simpleArchive (Politze & Krämer, 2017): combining necessary steps for creating a PID and archiving a data set in one interface. In addition to the mere preservation of the data bit stream, the context in which the data was created plays an important role. This context generally is captured by meta data. A minimum set of bibliographic meta data, as defined for example in the RADAR project (Kraft, et al., 2016), is required especially if publication of data is intended later in the data life cycle.

From an organizational point of view, it is further necessary to include meta data about the context of the person who created the data set, such as membership of institutions or chairs. With this information, it is possible to control which data records are visible for whom, especially in the case of (yet) unpublished data. This is especially essential as durations of data archival often exceed durations of employment and thus enables service providers to identify long-term responsibilities for archived data sets. This process again integrates with the PID workflow such that meta data and data sets are also linked via the PID. First, the PID is an URN to identify the data set in the meta data. In addition, an attribute "META_URL" is created in the PID to link the meta data record and to make it retrievable.

Based on the technologies presented, various processes can now be defined that support a decentralized data management system. Initially, a corresponding process was implemented at RWTH Aachen University.

2. REFERENCES

Curdt, C., Hoffmeister, D., Jekel, C., Udelhoven, K., Waldhoff, G., & Bareth, G. (2016). Implementation of a centralized data management system for the CRC Transregio 32 'Patterns in Soil-Vegetation-Atmosphere-Systems'. In C. Curdt, & C. Wilmes, *Proceedings of the 2nd Data Management Workshop* (pp. 27–33). Cologne, Germany.

Dreyer, M., & Vollmer, A. (2016). An Integral Approach to Support Research Data Management at the Humboldt-Universität zu Berlin. In Y. Salmatzidis. Thessaloniki, Greece.

Dumas, M., van der Aalst, W., & ter Hofstede, A. (Eds.). (2005). *Process-aware information systems.* Hoboken, NJ: Wiley.

Eifert, T., & Bunsen, G. (2013). Grundlagen und Entwicklung von Identity Management an der RWTH Aachen. *PIK - Praxis der Informationsverarbeitung und Kommunikation, 36*(2).

Kraft, A., Razum, M., Potthoff, J., Porzel, A., Engel, T., Lange, F., . . . Furtado, F. (2016). The RADAR Project - A Service for Research Data Archival and Publication. *ISPRS International Journal of Geo-Information, 5*(3), p. 28.

Politze, M., & Krämer, F. (2017). simpleArchive – Making an Archive Accessible to the User. In R. Vogl. Münster, Germany.