On the Decentralization of IT Infrastructures for Research Data Management

Marius Politze[1], Thomas Eifert[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, eifert@itc.rwth-aachen.de

Keywords
eScience, decentralized infrastructure, research data management, research process

1. ABSTRACT

National and international initiatives and the spread of data driven algorithms, recently put Research data management to more attention. Researchers can already choose from a variety of services; however, tasks like publishing, keeping public records or long-term storage often remain disjointed in researchers' workflows. While viewed as a university wide and generic task, the benefit of data management for researchers is likely indirect. Hence, the integration into research processes is a central challenge when establishing IT support for a data management system. From the IT service providers perspective this marks only a fine line between centralized generic services and specialized individual ones. Considerations between individualization, efficiency, personalization and sustainability are becoming more important, especially when it comes to publication and long-term storage of research data or for researchers' processes relying on central components.

We therefore strive to integrate individual research processes with central infrastructures. The realization of such a decentralized data life cycle depends on a shared set of services: digital identifiers like DOI and/or handle, meta data standards and formats, technologies for data preservation and interfaces for research process integration. This set of technologies allowed defining and implementing various processes within the data management system at RWTH Aachen University.

2. A NEED FOR DECENTRALIZED IT INFRASTRUCTURES

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, as shown by the repositories created within as part of TR CRC 32 (Curdt, et al., 2016) or LIFE (Kirsten, Kiel, Wagner, Rühle, & Löffler, 2017). Tailoring an IT infrastructure to the specific requirements of certain disciplines thus is crucial for their acceptance and short-term success. If, however, data management is viewed as a purely institutional, university wide and generic task, the benefit for researchers is often only indirect and therefore less obvious, resulting in reluctant usage (Eifert, Schilling, Bauer, Krämer, & Lopez, 2017). The integration into the processes of the researchers represents a central challenge when establishing IT support for a data management system.

RWTH Aachen University as well as many other universities have a highly decentralized technical infrastructure. This is particularly important when data is exchanged between systems that have to be connected (Haim, 2017) (Eifert & Bunsen, 2013). IT services supporting research data management originate not only 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 research groups and make these available to others within the framework of collaborations (Curdt, et al., 2016). The result is a heterogeneous, decentralized IT system landscape.

Central processes and offers that are so deeply integrated into the daily work processes of researchers, such as research data management, must not only consider this decentralization, but also actively deal with it in order to be attractive for the users and to be able to offer benefit. With the dawn of more decentralized services, their sustainability is becoming more important: Especially for the

publication and long-term storage of research data. In organizational terms, this is often the responsibility of the central service providers. At the latest 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 a data life cycle as shown in Figure 1, the systems used must enable this transfer as seamlessly as possible.

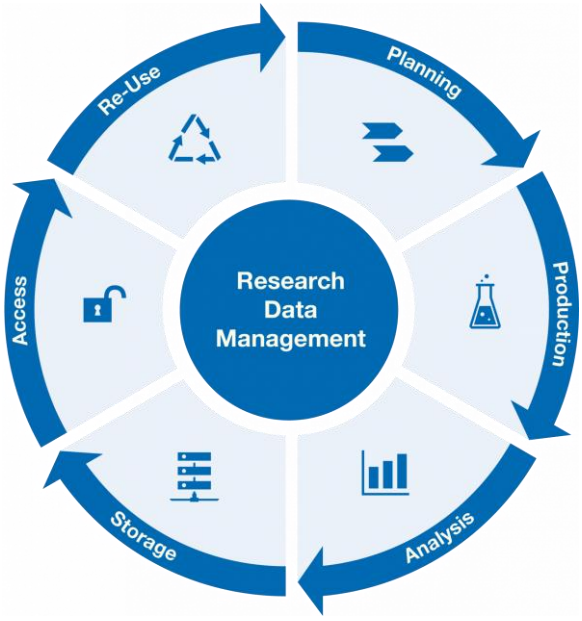

Figure 1: Research data life cycle at RWTH Aachen University

3. CONNECTING DECENTRALIZED INFRASTRUCTURES

In the field of data management, retention periods of ten years or more are quite common. In projects that deal with the digitalization of the "classical" industries the required retention periods equal that of these industries' objects, counting in decades. For the IT infrastructure, however, this period exceeds the lifetime of most IT equipment and this is quite a challenge: Typical maintenance contracts for server hardware, which run for five years but also increasingly fast-moving software life cycles, ensure that data repositories have to be migrated several times between successive systems during their retention time. In addition, there is often a dependency on the manufacturer to support the migration from one version to another. Services that support data management must therefore be designed so that the preserved data can be migrated. Especially with long-term IT services, it is therefore inevitable to plan how systems can be replaced or how and when the service can be switched off completely. This so-called exit scenario must therefore be part of the consideration right from the start.

One way to deal with this problem is to define technology-independent and process-oriented interfaces. Instead of specific formats and protocols, an independent and as open a specification as possible is used that is oriented to the processes to be supported. This abstract intermediate layer then takes over the translation between the processes of the researchers and the implementation of the 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).

The necessary modelling of processes requires interdisciplinary cooperation between technical and organizational service providers and (future) user groups. The latter are often not used to being included in the development and, from their point of view, fear a considerable additional effort. However, this is rewarded by the fact that appropriately modeled systems are better suited to the researchers' processes and that otherwise often necessary individual adaptations are omitted.

Translating between discipline specific requirements and IT services that can satisfy some of them is becoming steadily more important to allow scaling of services and infrastructures. However, it is not only a problem to connect process and service levels but also to interconnect services offering the

same functionality, a core problem within distributed service architectures (Wiederhold & Genesereth, 1997). The need for mediator services is becoming obvious by looking at the example of different services offering “cloud storage” like Google Drive, Dropbox, Box, or Owncloud Instances. Their common feature is storing files, but interfaces and especially APIs diverge in structure and functionality. Especially their increasing number and fast-moving development require new methods to construct mediators as interoperability layers to integrate services and research processes (Bennaceur & Issarny, 2015).

In addition to simplifying migration and exit scenarios, this procedure offers the option of integrating existing, heterogeneous system landscapes. As a result, technical boundaries between individual systems blur or even disappear completely. This bundling of existing services ultimately creates the desired benefit for researchers (Juling, 2009).

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.

3.1. Data Identification

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) (Kálmán, Kurzawe, & Schwardmann, 2012). In contrast to the DOI (Digital Object Identifier) system (Paskin, 2009), ePIC allows a very flexible and, above all, lightweight service to create PIDs. This enables minting of PIDs for data records as early as possible in the data life cycle, ideally directly when generating the data.

A so-called landing page is required to enable the resolution of the PID via a handle resolver. The URL to this page is stored in the attribute "URL" of the PID. For the implementation at RWTH Aachen University, a generic landing page is used to make it as easy as possible for researchers to create PIDs. This generic page offers the possibility to contact the researcher who created the PID, but does not provide any further information. In addition, the researcher can use the attribute "DATA_URL" to maintain a local reference to the data. This URL, however, is not necessarily publicly resolvable but serves only as an internal reference for the researchers to identify their data internally.

The PIDs created in this way 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.

3.2. Archiving and Preservation

In addition to the description of the research data, long-term storage plays an important role at the long tail of the data life cycle. Data is no longer changed and the access frequency decreases. Until now, tape archive is particularly suitable for data that is not published but should be kept in case it is to be reused. Nowadays, capacity-optimized incarnations of object storage are suitable alternatives. The current, still tape based, central archive of the RWTH Aachen comprises roughly 1PB of data in December 2018, of which about 1/3 was delivered in 2018, which shows the increasing importance of archival processes as part of research data management.

In order to make these technologies more accessible to researchers, a simplified workflow was established at the RWTH Aachen with simpleArchive (Politze & Krämer, 2017). This combines the necessary steps for creating a PID and archiving a data set in one interface. simpleArchive additionally integrates with the previously discussed PID workflow and hence also uses the attribute "DATA_URL" to store a reference to the place of archiving. Together with the account, information of the archive a contact is also possible.

3.3. Meta data

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.

Meta data can be further enriched with subject-specific information. In order to achieve the greatest possible flexibility at this point, the implementation at RWTH Aachen uses the Resource Description Format (RDF) is used for the representation of the meta data. An additional tool allows researchers to generate RDF compliant meta data based on preselected vocabularies and meta data schemas using a web form. A simplified REST interface also enables (partial) automation and adaptation to individual workflows of a research group (Politze, Bensberg, & Müller, 2019). The information generated in this way is stored in a graph database, Virtuoso, where it can be searched in the sense of a knowledge graph.

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 used as URN to identify the data set in the generated RDF. In addition, an attribute "META_URL" is created in the PID to link the meta data record and make it retrievable using the PID information. An even more close connection of PIDs and meta data is proposed by PID Information Types and their registry (Weigel & DiLauro, 2013). In addition to the data set itself, PID Information Types assign PIDs to individual meta data properties that can be resolved using the registry. PID attributes within the data sets' PID then use the meta data properties' PIDs to attach meta data to the data set. While the definition of PIDs, and therefore implicitly URLs, is compatible to the linked data model used by RDF and reflects the common practice to denote meta data standards and their fields. However, key value pairs of PID Information Types lack overall expressiveness and standardization.

Independently of the representation of meta data, particularly two points of the research data life cycle are critical: the actual 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 in time lie apart, the more time-consuming it is to obtain missing information about the data set concerned.

It is therefore necessary to collect this information as early 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. The actual data set does not have to leave the system used locally by the researcher. Only the existence of the data set, bibliographic and necessary discipline specific meta data required later must be recorded. The data record is thus centrally verified, if possible at the time of its creation. Each data record receives a persistent ID and can thus be identified independently of the actual storage location. In order to generate both long-term and direct benefit, the data set can be enriched with further, subject-specific meta data to enable searchability and retrievability on the basis of further research questions (Kirsten, Kiel, Wagner, Rühle, & Löffler, 2017). An appropriately established data repository could thus form the basis for a scientific "knowledge graph" (Decker, 2017) (Auer & Mann, 2018).

3.4. Research Process Integration

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:

The starting point is the addition of a text publication to the university bibliography. At the end of this step, researchers are asked to publish the previously decentralized data on which the publication is based or to archive it for internal reuse with the central service.

In general, researchers should prefer commonly known and, if possible, subject-specific repositories for the publication of research data. Publication in an institutional repository should only be chosen if no other repository is found. In any case, the dataset should receive a DOI that can be used to link the

text publication with the dataset. This type of link can also be used if only the meta data of the dataset is publicly available. In both cases, mandatory data such as author, title and year of publication are published. This further underlines the importance of bibliographic meta data to be present to handle data publications later in the research data life cycle.

However, it is sometimes required by researchers to publish any information about the data. In order to achieve an unambiguous link between text publication and dataset, the corresponding PID is used. In order to make this process as simple as possible for the researchers, the RWTH Publications repository and the simpleArchive Workflow are connected and information on archived data is automatically exchanged.

4. FOCUS FOR FUTURE IT INFRASTRUCTURE DEVELOPMENT

As data management tasks are becoming inevitable within research processes, future IT services and infrastructures need to put more emphasis on this topic. The necessity to bring together well-standardized and scalable IT infrastructures with highly specialized and individual research workflows poses additional requirements towards future IT service development. Integration of different building blocks is key to solve these future challenges.

IT services are no longer isolated but are becoming a building block within the researchers' workflows as illustrated by **Fehler! Verweisquelle konnte nicht gefunden werden.**. A basic requirement towards current future services hence is for open APIs. Service providers unfortunately still connote the term "open" with "publicly accessible" or, even worse, "insecure". While the former is likely empiric, the latter simply is a misconception. From a workflow and integration perspective, the concept of open APIs relies on publicly available descriptions of an IT service and its stable APIs. Both, the APIs and its description, should further be available through a standardized, and platform independent protocol. "Publicly" in this case should furthermore be understood as available to the customers. This further implies that open APIs can (and probably should) be secured to be accessed by validated customers only.

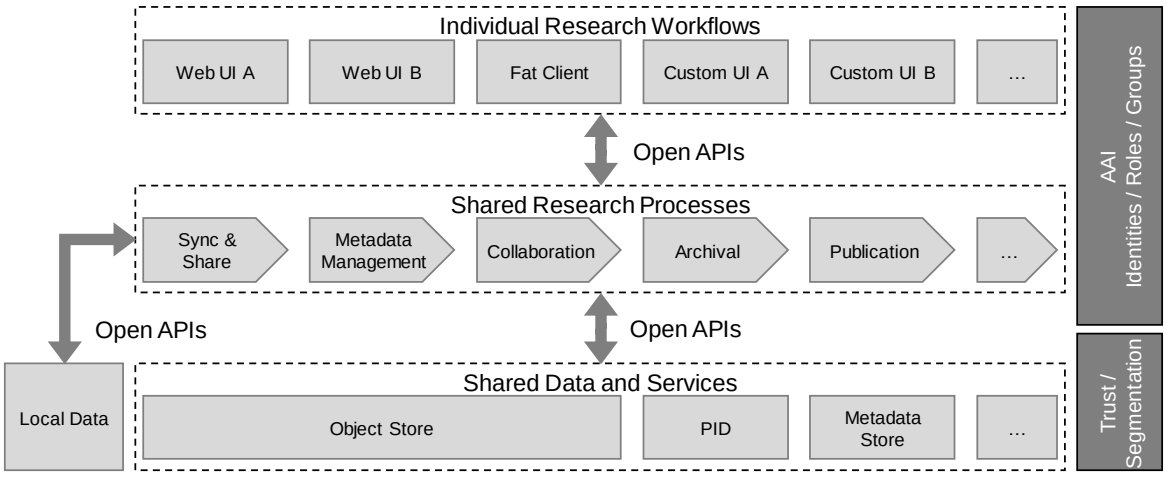

Figure 2: Schematic integration of shared and local services into process supporting layers

However, open APIs becoming de-facto industry standards are becoming an issue when rebuilt by competitors as shown by the prominent Oracle-Google case (Finley, 2016). API specifications should therefore have a clear licensing of their own and independent from the implementing software. Hence, future IT services should furthermore encourage transitioning not only to open APIs but also to open standards for APIs: API specifications that are (publicly) available and that are licensed to be recreated by other systems.

This API oriented approach also helps the scientific users to integrate several independent functional units into their individual / their team's workflow. From an abstract perspective this is a type of high level programming or automation that becomes common with increase of digital literacy.

Defining open APIs and/or standards, however, only takes one side of the coin. IT service providers need to adhere to them rather than defining their entirely new or deviating and incompatible versions. Standardization may be a slow process and may therefore hinder innovation. An open standard also

should not be considered fundamental truth. Future IT systems should be able to translate between competing standards and offer APIs for different purposes or workflows. This context of an IT service is becoming more important as service providers are becoming solution providers and IT services are becoming more expedient than essential parts of a solution. Gradually IT services are no longer worthy ends in themselves. Reducing IT services to the (open) standards they comply to thus makes it easier to align requirements, service portfolios, ongoing maintenance and service replacement.

One aspect when dealing with decentralized processes is the vanishing validity of assumptions concerning the client devices. Where formerly the usual PC was centrally managed and part of a system environment, the digital literacy of the users leads to a demand of individually configured devices. From a process perspective this is indistinguishable from a “BYOD” approach: a device fully controlled and configured by the user. The ownership structure is unimportant here. Considering research workflows, these devices are furthermore not limited to computers, smartphones or tablets only but may include IOT-Sensors, measurement devices or other “smart” lab equipment. Under these conditions neither a certain software set nor a certain OS configuration can be relied on. In particular, the client offers no reliable user authentication or any other kind of rich user interfaces. In such a scenario, the IT infrastructure has to provide all these components, from secure Authentication and Authorization Infrastructure (AAI) over reliable data storage to a suitable software collection and an architecture to connect these components together.

5. CONCLUSION

The digitalization of research workflows is tackled from two sides: Researchers adopt and improve their existing workflows and IT service providers that offer ever enhanced IT services. Neither implementing individual workflows nor offering generic services only will be sufficient to successfully shape the future of research data management. Consistent workflows throughout the research data life cycle that incorporate discipline specific characteristics need to be supported by sustainable and scalable services. Services need to coexist with each other and adapt to standardized and individual workflows.

The proposed components span a decentralized data management system and try to incorporate such existing workflows of researchers as far as possible. Researchers can choose from various offers for the description, archiving, verification and publication of data and build upon these processes to refine their own data management workflows. Data identifiers like ePIC PIDs or DOIs are key to identify and trace data across its life cycle. A discipline specific Meta data management then is key to manage data on an institutional level and to make data accessible for future generations of researchers. While, the implemented process for linking text publications and data is strongly oriented towards the needs of the researchers, it is only situated at the end of the data life cycle. Further, equally integrated offers for a complete support are necessary. However, the workflows show a way with which a data management system can be set up without dependency on a central institutional data repository.

It is clear that data management systems must start as early as possible in the data life cycle. Early identification and description play an important role. Despite all the necessity, many researchers fear the additional effort. Future IT Systems need to help individual researchers by directly integrating into their local workflows. With ever increasing digital literacy, research processes and workflows are becoming more complex and integrate more individual client devices. This can only be supported effectively if future IT Systems offer open APIs and adhere to open standards to integrate existing services or to be integrated into an existing workflows.

6. REFERENCES

Auer, S., & Mann, S. (2018). Toward an Open Knowledge Research Graph. *The Serials Librarian*, pp. 1–7.

Bennaceur, A., & Issarny, V. (2015). Automated Synthesis of Mediators to Support Component Interoperability. *IEEE Transactions on Software Engineering, 41*(3), pp. 221–240.

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.

Decker, S. (2017). *Rethinking access to Scientific Knowledge: Knowledge Graphs*. Retrieved 2 3, 2018, from https://www.linkedin.com/pulse/rethinking-scientific-knowledge-graphs-stefan-decker/

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).

Eifert, T., Schilling, U., Bauer, H.-J., Krämer, F., & Lopez, A. (2017). Infrastructure for Research Data Management as a Cross-University Project. In S. Yamamoto, *Human Interface and the Management of Information: Supporting Learning, Decision-Making and Collaboration* (Vol. 10274, pp. 493–502). Cham, Switzerland: Springer International Publishing.

Finley, K. (2016). *The Oracle-Google Case Will Decide the Future of Software*. Retrieved April 23, 2019, from Wired: https://www.wired.com/2016/05/oracle-google-case-will-decide-future-software/

Haim, M. (2017). Herausforderungen des Identity Management an Hochschulen - Problem Datenintegration. In P. Müller, B. Neumair, H. Reiser, & G. Dreo Rodosek, *10. DFN-Forum Kommunikationstechnologien* (pp. 63–74). Bonn, Germany: Köllen.

Juling, W. (2009). Vom Rechnernetz zu e-Science. *PIK - Praxis der Informationsverarbeitung und Kommunikation, 32*(1), pp. 33–36.

Kálmán, T., Kurzawe, D., & Schwardmann, U. (2012). European Persistent Identifier Consortium - PIDs für die Wissenschaft. In R. Altenhöner, & C. Oellers, *Langzeitarchivierung von Forschungsdaten* (pp. 151–164). Berlin, Germany: Scivero Verl.

Kirsten, T., Kiel, A., Wagner, J., Rühle, M., & Löffler, M. (2017). Selecting, Packaging, and Granting Access for Sharing Study Data. In M. Eibl, & M. Gaedke, *INFORMATIK 2017: Digitale Kulturen* (pp. 1381–1392). Bonn, Germany: Köllen.

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.

Paskin, N. (2009). Digital Object Identifier (DOI ®) System. In M. J. Bates, & M. N. Maack, *Encyclopedia of Library and Information Sciences, Third Edition* (Vol. 6, pp. 1586–1592). CRC Press.

Politze, M., & Krämer, F. (2017). simpleArchive – Making an Archive Accessible to the User. In *Proceedings of the 23rd EUNIS Congress* (pp. 121–123). Münster, Germany.

Politze, M., Bensberg, S., & Müller, M. S. (2019). Managing Discipline-Specific Metadata Within an Integrated Research Data Management System. In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019) - Volume 2* (pp. 253-260). Porto, Portugal: SCITEPRESS.

Weigel, T., & DiLauro, T. (2013). Separation of Concerns: PID Information Types and Domain Metadata. Lisbon, Portugal.

Wiederhold, G., & Genesereth, M. (1997). The conceptual basis for mediation services. *IEEE Expert, 12*(5), pp. 38–47.

7. AUTHORS' BIOGRAPHIES

Marius Politze is head of the group "Process and Application Development for Research" at the IT Center of RWTH Aachen University. His own 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.

Dr. Thomas Eifert is the CTO of the IT Center of RWTH Aachen University. In this role he evaluates new technologies for their suitability and is responsible for the future development. His research is on scalable IT solutions with special interest in research data management and e-learning. He studied physics and received his doctoral degree in solid state chemistry.