Introducing Research Data Management as a Service Suite at RWTH Aachen

Thomas Eifert[1], Stephan Muckel[2], Dominik Schmitz[3]

Abstract: Research Data Management (RDM) receives more and more attention as a core component of scientific work. This importance equally stems from the scientific work with ever-increasing amounts of data, the value of this data for subsequent use, and the formal requirements of funding agencies. While these requirements are widely accepted among the researchers, the individual acceptance depends on many factors. In this paper we describe the initial steps at RWTH Aachen University to implement RDM as a widely accepted service suite. Here, service means appropriate IT solutions as well as a comprehensive training and support. The steps include the preliminary work to get the top management's awareness and the dialog with the scientists to learn the practical requirements. We present the initial solutions and solution concepts derived so far. We explain how our local approach relates to broader external initiatives. It is especially important that all these concepts together encompass a comprehensive suite of support offerings, guidelines, and various IT service modules built or planned on the underlying IT infrastructure.

Keywords: Research Data Management, Service introduction

1 Introduction

Research data is the outcome as well as the foundation of scientific work. Researchers need an environment that enables them to work efficiently and securely with their research data (cf. [KE13, EU10]. Currently, research data management (RDM) at RWTH Aachen is not as well-developed at all levels as to support the research process in an optimal way. Moreover, (inter)national research funding institutions, such as the European Union program Horizon 2020, the German Research Foundation and the Federal Ministry of Education and Research as well as various publishers (e.g. NATURE[4]), are increasingly requiring scientists and scholars to plan and execute good data management practices. An obligation to archive produced data already exists by carrying out "good scientific practice" [RW11], some even ask for the publication of primary data, e.g. the Open Data Pilot of the EU[5].

To fulfill these growing requirements and thus make research data accessible and usable for subsequent research projects has become an inevitable objective for all scientific institutions. Thus, structures and processes have to be established to relieve the scientists

[1] RWTH Aachen University, IT Center, 52056 Aachen, eifert@itc.rwth-aachen.de
[2] RWTH Aachen University, Central Administration, 52056 Aachen, Stephan.muckel@zhv.rwth-aachen.de
[3] RWTH Aachen University, University Library, 52056 Aachen, d.schmitz@ub.rwth-aachen.de
[4] http://www.nature.com/srep/journal-policies/editorial-policies
[5] http://europa.eu/rapid/press-release_IP-13-1257_en.htm

from these tasks and let them focus their primary work. For this reason, there are many (inter-)nationally coordinated activities as well as activities at federal state and local levels to tackle these questions. This contribution presents the local approach to combine and integrate these activities into the local infrastructure to prepare the path for the steps to come.

Domain Model The well-known domain model [KE13] visualized in Figure 1 has proven to be a valuable means to a common understanding and basic structuring of research data management. It distinguishes four domains where researchers act: the private domain, the group domain, the persistent domain and the public domain enabling access and reuse. Research typically starts in the private domain where a researcher mainly works alone on his or her data. Depending on the circumstances e.g. the project situation the need arises to share the data within the group domain. Typically this forces to explicitly add metadata well-known to the individual researcher but if not provided prohibiting an understanding by collaborators. Once publications have been written or, at least, at the end of a project, good scientific practice asks for long-term storage of primary research data. Again, the need to capture more descriptive information rises since it cannot be ensured that people that have produced data will be available once the data need to be accessed. This is obvious for the public domain which can be chosen to be addressed by a researcher at any time. Using the data requires a thorough understanding that is helped by an encompassing documentation with comprehensive and as far as possible standardized metadata.

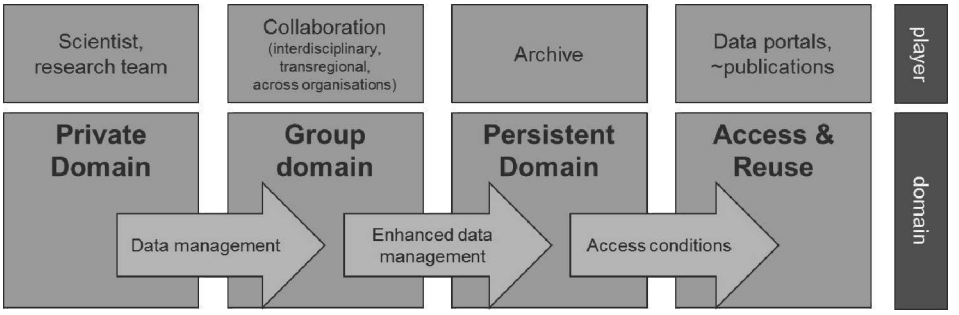

Figure 1: Domain model [KE13]

2 Initiating RDM: the Pre Project at RWTH

At the latest with the recommendations by the HRK, the German rectors' conference, [Ho14, Ho15] it has become obvious that universities as any other research institutes have to address research data management. When starting the project, we have observed similar approaches at many other places. Typically, three central institutions form the nucleus of a local project to establish research data management at an institutional level: the computing center for IT support and storage, the library for metadata and publishing

issues and the central administration in particular the part concerned with research funding. But a major concern is to get the researchers involved that are the addressees of such a service suite. Fortunately several national and international projects have been carried out or recently started, that address at least parts of these issues. We will refer to them throughout the rest of the paper.

2.1 Capture Current Situation

A typical approach to get the researchers involved is to simply ask them about their needs [Si14, Tr16]. The known surveys build on each other and the organizations are open to share their questionnaires which reduces the locally required efforts. Due to the already published survey results, we decided for the RWTH to start by a rather small sample of scientists that were interviewed in a structured manner. This analysis targeted at the creation of a starting point for planning an RDM infrastructure, to get into direct contact to communicate the aim of the project and to involve the affected scientists at the earliest point possible to address the "freedom of science" as well as fostering acceptance and awareness. Selecting the sample we took care of a broad coverage of criteria such as faculty, team size, age and gender, etc. The evaluation of the interviews later gave a good confidence that – at that level of detail – we got a reasonable coverage of all properties required.

The interviews were carried out along a self-developed guideline building on the experiences of larger surveys in particular [Tr16]. The invitation to participate in the interviews was sent by the University's top management, i.e. the rector. As a common result all scientists confirmed to already have methods for dealing with research data that accommodate their personal needs. Nonetheless, all observe a huge potential for improvement, in particular in regard to the improved capture of descriptive metadata that enable discoverability and reuse all alone. Yet they do not consider this a responsibility of their individual scientists or teams. Thus, while the survey did not bring up entirely new issues it ensured a mandate for a centrally organized service structure at RWTH.

2.2 Derived Requirements

The need for supportive services is seen consistently. But all interview partners stressed that any supportive service needs to be strictly user-oriented, intuitive to use, and easily be integrated into current research processes. In particular, we have to take into account the various discipline-specific repositories and virtual research environments that have been established over the last years. A CERN-related physicist has different processes and less local need than local laboratories with unique features. Thus, the functional and non-functional characteristics of a suitable service suite must be evaluated with respect to the local environment as well as national and international settings.

From the surveys and our own interviews we were able to derive requirements on future services for RDM support and non-functional qualities of these services that are essential

for acceptance. The different ranges of these qualities can directly be mapped onto the domain model (see Fig.1).

For the persistent domain, the predominant requirements are stability over time, an adequate metadata infrastructure to ensure discovery as well as scalability in size, since over time all research data will be stored here. Much more important than for the long term storage, the requirements to the UI for the daily “live” work, i.e., in the “private” and the “collaborative” domain, are very specific and very broad at the same time: The scientists expect traditional file system access as well as via web browser, through RDM systems or by cloud techniques, together with fast access to large resources and easy means to attach and link metadata to improve discovery, documentation, and reuse. We expect a further strong growth with respect to IT integration, for instance in the areas of experiment control and automated measurement systems, so the emerging solutions on this side will drive the way data as well as metadata is transferred to storage systems. Among these new scenarios are measurement systems which directly deliver raw data via cloud protocols and store the describing metadata in a database.

Common to all domains are the aspects of user-orientation, which means from the handling as well as from the functional perspective, and integration into existing workflows. On the other hand, the central institutions are not able to realize and maintain a homogeneous digitalized working space for each individual researcher. This will never scale. Accordingly, the challenge is to meet the right mixture of basic support and extensible infrastructure.

2.3 Support by University Top Management, Communication & Awareness

An important factor for success is that the university management – after internal discussion with all the stakeholder groups – starts by sending a clear, primarily internal, message that RDM will represent a key element of the university's ethos going forward. Publication of a "research data policy" in the form of strategic guidelines for RDM has often proven helpful in putting the subject of RDM at the top of the university agenda. Different University groups and committees are involved in the process to develop guidelines on research data management at RWTH. These guidelines, decided by rectorate in March 2016 [RW16], create awareness to take on responsibility for research data, to guarantee reproducibility, availability, and protection – including the protection of intellectual property –, and to provide functionalities concerning organizational, description/metadata, archiving, and reusability related aspects.

3 Current Project and Intended Solutions

The RWTH rectorate has taken responsibility for the implementation of HRK Recommendation of 2014-05-13 on the topic "Management of research data – a key strategic challenge for university management" [Ho14] and following the

recommendation of the RWTH Aachen CIO Advisory Board has decided to implement the following measures: establishing a holistic consultancy and support offer, a specification of resources (costs, persons), suitable awareness and acceptance activities, competence building teaching offers not only for researchers but also already for master students, strong need for prolific cooperation partners, stepwise introduction of suitable IT support, introduction of a suitable RWTH policy on research data management, and the establishment of a holistic IT support for the entire research process for at least three institutes at RWTH until mid of 2017.

3.1 Structure of the Project

To fulfill the university top management's tasks five modules have been established within the project. The module "awareness and acceptance" takes care of the very important communication activities. As also stated by the German Rectors' Conference [Ho15] this is a pivotal and major concern within a successful RDM project. As a specific part the setup of a broadly accepted "RDM policy" is established as a separate module. Further on, another major concern that is also very much related to awareness is the module on "consultancy and RDM teaching". It forms one of the central corner stones of the project. But without suitable IT support nobody will be able to establish a tangible research data management in our digital world, thus the equally urgent need for a module on "technical support". Yet, this module is not intended to resolve all problems by developing all solutions again locally. Thus, to achieve sustainable financing of RDM activities, a fifth module explicitly addresses "cooperations, partners and projects", thus building on initiatives such as the Research Data Alliance (RDA)[6] or the DINI/nestor working group on research data management[7].

In the following sub chapters we will present first glimpses on the intended solutions on consultancy and teaching offers as well as technical IT support that is embedded in the existing as well as externally developed IT infrastructure.

3.2 Consulting Services and Teaching

Already since 2014 the university library has established an introductory course for PhD students covering all basic aspects of research data management. It has been adapted to the feedback and is nowadays offered twice per semester. This introductory course serves as the foundation to elaborate an encompassing course offer that details out particular topics such as personal (meta)data management, data management plans, publishing and discovering data, collaboration, and archiving.

Regarding consulting services we currently follow two trails. On the one hand we have chosen a small number of dedicated institutes as scientific partners in the project to get

[6] http://rd-alliance.org
[7] http://www.forschungsdaten.org/index.php/AG_Forschungsdaten

real world experiences with the concrete problems researchers face during their daily work. The intention is to develop solutions for and with them during the project phase with the aim to derive more generic solution patterns that can be applied in similar institutes. The second trail is that we positioned the established Service Desk run by the IT Center as the single point of contact. The people from service desk are informed where to delegate questions as 2nd level support thereby hiding the complexity of separate institutions within the university supporting research data management (IT Center, library, central administration). For more complex problems, we plan a 3rd level that in fact realizes mini projects combining suitable forces from central institutions and the affected institute. The main purpose of this integrated set up is, to be able to scale. As more and more questions arrive at the service desk, the more repeated questions can be expected that can with growing knowledge then be answered already in the 1st level support while currently nearly all questions will be forwarded to the 2nd level.

In regard to the content of consultancy and teaching, we assist scientists with the creation of data management plans and advise them on their realization as well as on how to handle research data in their project. In particular, we help scientists to find a suitable metadata model for depicting their data or develop an appropriate solution together with them. Directories of standards and ontologies such as the RDA Metadata Standards Directory Working Group[8] typically serve as the starting point. In many cases domain specific approaches exist and are well suited to address researchers' needs.

3.3 Conceptualization of the Technical Infrastructure

The conceptualization of the technical infrastructure is driven by the observation, that a fully integrated and fully individualized monolithic digitalized working space for each individual researcher might be optimal in regard to meeting the user requirements (see Chapter 2.2) but it is unrealistic to be ever achieved by a single local institution. The span of domain- and discipline-specific requirements is simply too large. Accordingly, the solution concept foresees modularization and scaling options as the cornerstones of the technical infrastructure. In addition, only by reconsidering existing services – locally as well as provided by external partners or services at national or international level – and incrementally adding new services the introduction of a holistic service suite for research data management becomes feasible.

We basically distinguish between two aspects: for one there must be enough place to safely store the ever increasing amount of research data and on top of that we must provide suitable services that allow to work effectively and efficiently with the data.

[8] https://www.rd-alliance.org/groups/metadata-standards-directory-working-group.html

3.4 Basic Needs and Services

When figuring out where to start it has become obvious that despite the fact that the earlier encompassing metadata to research data are established the easier any later step and activity can be performed, the corresponding change in mind of researchers cannot be expected to take place too soon – many communication, consultancy and teaching activities will be needed to pave the ground. Yet the urgent need to improve the current situation has already been becoming obvious from our small local survey. In particular, researchers must be relieved from the burden of ensuring the most basic requirements of good scientific practice, mainly the requirement to ensure safe storage of research data for at least ten years after research completion, i.e. the persistent domain and optionally the access and reuse domain. Following the concept of modularization, at the same time the established IT system structure should be open and supportive to any more advanced approach to research data management also addressing the other earlier and more complex domains, the private and the group domain.

The university library has taken over the duty to cover the access and reuse domain. They provide researchers with information on where to publish their subject specific data (e.g. specific repositories as they can be found at re3data [Pa15]) also complying with funder requirements e.g. by EU. This also includes the opportunity to publish data at the already established repository[9] at RWTH Aachen if no suitable subject specific repository exists.

To address the need of a non-publishing but archiving facility, we have looked at several typical processes (data to a dissertation, data to a project, data to a publication) and have extracted a minimal process for data archiving, shown in Figure 2.

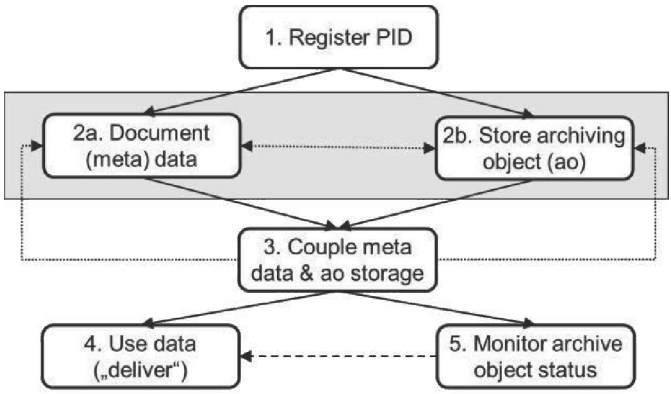

Figure 2: Basic Data Archiving Workflow

As a result, this basic workflow asks for three minimal IT services that need to be established.

[9] http://publications.rwth-aachen.de

1. An independent, highly available and performant persistent identifier (PID) service that can be addressed at any time by any researcher for as many objects as needed.

2. A storage component amenable to the needs of the researcher. While central basic solutions are offered – in particular access to the central tape archiving infrastructure –, due to the heterogeneity of research at RWTH we do not expect a single solution to be used. Thus at the PID service, any location can be registered.

3. To enable suitable subject specific and individual approaches to the description of resources, we similarly plan to register a link to additional metadata within the PID system. This builds on the RADAR schema[10] as the most basic set of metadata.

While the high amount of industrial cooperations at RWTH Aachen as well as the limitations of RADAR [Kr16] to store data related to persons[11] ask for a local solution, we consider it important to build the local solution in accordance with this national approach. In regard to the PID service, we will indeed be able to make use of the EPIC infrastructure provided by GWDG Göttingen within the EU project EUDAT[12]. Other examples for current and future service modules are Web tools for metadata libraries, version control systems and others. One particular service already in place is a university-wide identity management system [EB13] that allows account provisioning attached to person lifecycle management. With respect to RDM the knowledge about the creator of scientific data is as important as access management by identity based roles.

By this basic but open infrastructure we expect to address the urgent current needs as well as to be open enough to subject and institute specific solutions that the individual researchers plan for. In particular, more advanced data processing pipelines can be established that produce data, register them with PIDs and refer to them via the PIDs in later analyses steps as it has been proposed in the notion of the "data fabric" [Wi15].

3.5 Derived Long-Term Concept for the Storage Layer

Quantitative characteristics must be derived from the information collected in the interviews and the current storage usage for now as well as for the foreseeable development in order to create reasonably sized support and operating structures as well as technical equipment. From the storage layer perspective, we derived a 3-tier architecture to support the needs of a future RDM system as well as the currently available service modules.

The lowermost tier corresponds directly to the persistent domain and is realized by a high-capacity system (based on tape technology). The capacity is oriented at the total

[10] https://www.radar-projekt.org/download/attachments/3178555/AP3_RADAR_Metadata-Kernel__engl_v02.pdf?version=1&modificationDate=1415809915000&api=v2
[11] https://www.radar-projekt.org/display/RD/FAQ
[12] http://www.pidconsortium.eu/

amount of data that has to be kept in the context of RDM. Since there is little live access to the data stored here this tier could easily be shared with other universities when the privacy protection issues can be solved.

The mid-tier has to support all access methods depicted in Chap. 2.2. Capacity wise, we see this tier to accommodate all running projects, so it has to keep all data generated during typical project or thesis durations. It is also the place where collaborative access will be realized, either in the context of the collaborative domain as in the Access-and-reuse-domain. The most promising technologies for implementing this tier have high scaling characteristics in terms of size and speed as we see it in current disc based so-called object storage solutions.

The uppermost tier has to accommodate the performance and functionality of highly interactive work as it is carried out in the private and the group domain. For convenient user access, this tier has to realize a multitude of current and future access protocols, from fileserver to web/cloud methods. With respect to the data, this tier holds a sort of a working set of data. It is also the tier where the use-related requirements, in particular the intuitive UI and the ability to integrate with well-established processes within the scientific departments, are most important. Even more, a bad match to these “soft” requirements induces a high risk of bad acceptance by the scientists.

4 Outlook

Based on the results we gained so far we are currently developing our set of existing services in the direction of RDM. During this process we identify gaps and start projects to close these. On the other hand, now that we have a concept of how to accommodate the storage needs of RDM we cast this concept in a series of cooperative proposals in order to get funding support either for the necessary work as well as for the large scale storage components.

Acknowledgements

The depicted work has been carried out by a joint project group of the University’s Library (U. Eich, S. v.d. Ropp, D. Schmitz, U. Trautwein-Bruns), IT center (B. Magrean, I. Hengstebeck, Th. Eifert, F. Krämer, A.-K. Wluka), and central administration (J. Dautzenberg, A. Haverbusch, S. Muckel) led by Prof. M. S. Müller and E. Müller.

References

[EU10] European Union: Riding the Wave. How Europe can gain from the rising tide of scientific data. Online: http://ec.europa.eu/information_society/newsroom/cf/

document.cfm?action=display&doc_id=707, Last Access: 2016-01-11, 2010.

[Ho14] Hochschulrektorenkonferenz: Management of research data – a key strategic challenge for university management. Online: http://www.hrk.de/positionen/gesamtliste-beschluesse/position/convention/management-von-forschungsdaten-eine-zentrale-strategische-herausforderung-fuer-hochschulleitungen/, Last access: 2016-01-12, 2014.

[Ho15] Hochschulrektorenkonferenz: How university management can guide the development of research data management. Orientation paths, options for action and scenarios. Online: http://www.hrk.de/positionen/gesamtliste-beschluesse/position/convention/wiehochschulleitungen-die-entwicklung-des-forschungsdatenmanagements-steuern-koennen-orientierung/, Last access: 2016-01-12, 2015.

[KE13] Klar, J., Enke, H.: Rahmenbedingungen einer disziplinübergreifenden Forschungsdateninfrastruktur, Report Organisation und Struktur. DOI: 10.2312/RADIESCHEN_005, 2013.

[Kr16] Kraft, A. et.al.: The RADAR Project—A Service for Research Data Archival and Publication. ISPRS Int. J. Geo-Inf. 5, 28. DOI: 10.3390/ijgi5030028, 2016.

[Pa15] Pampel, H. et.al.: Stand und Perspektive des globalen Verzeichnisses von Forschungsdaten-Repositorien re3data.org. In (Müller, P. et.al., eds): 8. DFN-Forum Kommunikationstechnologien: Beiträge der Fachtagung 08.06-09.06.2015 in Lübeck (GI-Edition: lecture notes in informatics; 243, pp. 13–22). Gesellschaft für Informatik, Bonn, 2015.

[RW11] RWTH Aachen University: Grundsätze zur Sicherung guter wissenschaftlicher Praxis der Rheinisch-Westfälischen Technischen Hochschule Aachen. Amtliche Bekanntmachung vom 11.01.2011. Online: http://www.rwth-aachen.de/global/show_document.asp?id=aaaaaaaaaaoyxb, Last access: 2016-01-12. 2011.

[RW16] RWTH Aachen University: Leitlinien zum Forschungsdatenmanagement für die RWTH Aachen. Rektoratsbeschluss vom 08.03.2016. Online: http://www.rwth-aachen.de/global/show_document.asp?id=aaaaaaaaaaqwpfe&download=1, Last access: 2016-03-10, 2016.

[Si14] Simukovic, E. et.al.: Was sind Ihre Forschungsdaten? Interviews mit Wissenschaftlern der Humboldt-Universität zu Berlin. Bericht, Version 1.0. Online: urn:nbn:de:kobv:11-100224755, Last access: 2016-01-12, 2014.

[Tr16] Tristram, F. et.al.: Öffentlicher Abschlussbericht von bwFDM Communities – Wissenschaftliches Datenmanagement an den Universitäten Baden Württembergs. Online: http://bwfdm.scc.kit.edu/downloads/Abschlussbericht.pdf, Last access: 2016-04-08, 2016.

[EB13] Eifert, T., Bunsen, G.: Grundlagen und Entwicklung von Identity Management an der RWTH Aachen. PIK - Praxis der Informationsverarbeitung und Kommunikation. Band 36, Heft 2, Seiten 109–116, DOI: 10.1515/pik-2012-0053, 2013.

[Wi15] Wittenburg, P.: Data Foundation & Terminology WG. Data Fabric IG. Presentation at the RDA-DE-DINI Workshop “Aktuelle Resultate der Research Data Alliance (RDA) und deren zukünftige Bedeutung” 2015-05-28/29 in Karlsruhe, Germany. Online: http://www.forschungsdaten.org/images/f/fc/RDA-DE-2015_Wittenburg_Peter_III.pdf Last access: 2016-01-12, 2015.