An Integral Approach to Support Research Data Management at the Humboldt-Universität zu Berlin

Malte Dreyer[1], Andreas Vollmer[1]

[1]Computer and Media Service, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany, malte.dreyer@cms.hu-berlin.de, andreas.vollmer@cms.hu-berlin.de

1. Abstract

To address the increasing requirements on research data management, the Humboldt-Universität zu Berlin (HU) developed and established an integrated approach to provide research data support and services to scholars, students and administrative staff. This paper will discuss the type of requirements from external stakeholders, as well as the needs expressed and identified from within the HU. The derived concept will be explained in detail and examples for specific activities will be given.

2. Introduction

Results or products of research are not just solely publications (Piwowar, 2013), but also raw data, data aggregations, visualizations, software or even complex infrastructures. Compared to traditional research products like papers and books, digital artifacts change and eventually vanish much faster when not taken care for in a proper way (Vines et al., 2014). Many common publication formats have developed throughout the last centuries of traditional publishing and still influence how publications look like today (Gross, Harmon, & Reidy, 2002). Common standards for handling and enriching digital research data are not yet widely established. Data formats are far more heterogeneous (Arms & Fleischhauer, 2005), there are just few stable formats which last more than a decade, processing data is mostly individual (Hey, Tansley, Tolle, & others, 2009). Even long term storage and preservation, besides being costly, is not a commodity yet, like it is for storing books or journals (Palaiologk, Economides, Tjalsma, & Sesink, 2012). Data is being produced in vast amounts and unlike for traditional publications, individual data are often not enriched by metadata, categorized or described by abstracts by the libraries (Wessels et al., 2014). Hence, proper data management and data literacy is an increasing requirement on research processes. In the context of research transparency, fidelity and veracity as well as good scientific practice, more and more research funders mandate certain methods of research data management or at least require a documented reflection on the intended data handling processes (Keralis, Stark, Halbert, & Moen, 2013). Data literacy or “research data literacy” (Schneider, 2013) however can't just be seen as a requirement from funders about research transparency. Proper data handling, well reflected views on data quality and fidelity and related software also enables for extended capabilities to re-use data and to develop more sustainable software systems.

Among other aspects, this results in research data being a far more ephemeral and fragile format than their paper pendants (Warner & Buschman, 2005). What is more, addressing the open science paradigm of sharing research data even extends the scope of complexity to specific scholarly cultures (Borgman, 2012).

As scientific disciplines or domains differ considerably in requirements, technological preferences, standards and collaboration schemes, global activities like the Research Data Alliance (RDA) are a crucial community activity to work on the plethora of issues involved and explore specific problem domains, provide guidelines or document existing diversity (Berman, Wilkinson, & Wood, 2014). Exploiting these recommendations and results enables to adjust objectives and focus on more stable task domains.

It still remains a complex task to foster and establish research data management and research data literacy within a bigger higher education institution like HU.

3. Identifying requirements

The Computer and Media Service (CMS) as the data center of the HU is active in this field for many years. By establishing a position for research data management coordination, the HU was able to invest capacity into improving the mutual communication on diverse scientific processes and central infrastructure.

To evaluate the current status, best practices, requirements and gaps concerning research data and related methods at HU, the data center conducted a survey in January 2013. With around 500 responses the results helped to shape a research data strategy (Schirmbacher, Kindling, & Simukovic, 2014). Main requirements from scholars as well as service gaps could be identified. The following illustration depicts the strongest demands expressed by the scholars:

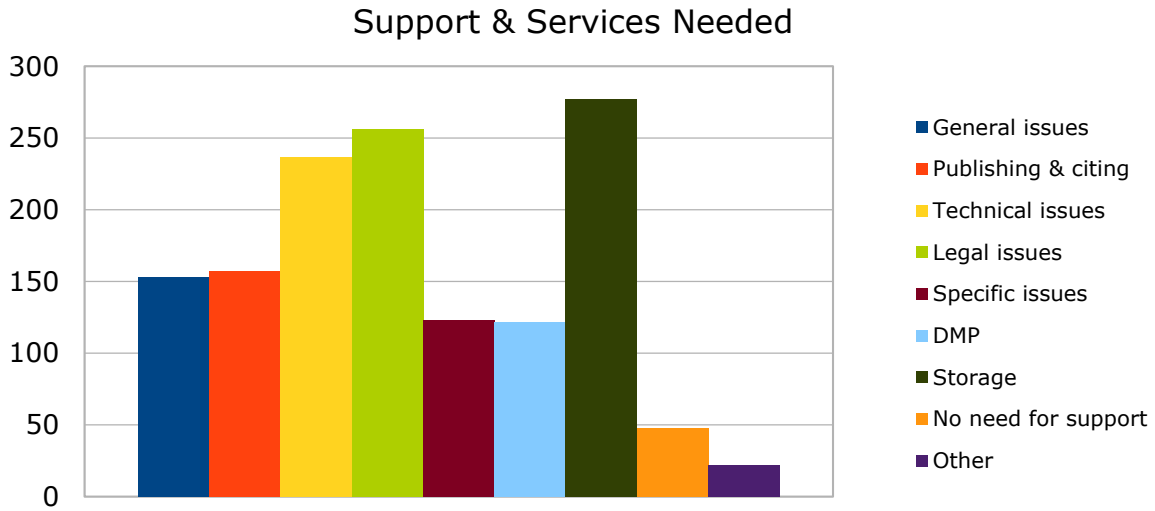

Figure 1: Expressed needs for support (Schirmbacher, Kindling, & Simukovic, 2013)

As a result of this survey, the strongest requirement is in the area of storage, followed by legal and technical issues. These issues were addressed first. Storage surprisingly got the highest number of responses, although the data center offers a broad set of storage services for many purposes, still it appears to be insufficient. This contradiction finally resulted in a new storage service implemented.

As a next step after the survey, a research data policy was designed and discussed with all stakeholders involved. It was adopted by the senate of Humboldt University July 8th 2014.

The policy defines basic principles in the context of safeguarding good scientific practices like documenting the data lifecycle or proper preparation for long-term archiving ("Humboldt-Universität zu Berlin Research Data Management Policy," 2014).

4. Building blocks

Based on these preparations, in 2015 a new concept was designed as an integrated approach to foster, facilitate and support research data management at HU. The main components are:

- **Community Building** to identify most active stakeholders and establish working groups.
- **Building tools and services** to address needs that have been expressed frequently, like easy to use storage services, DOI assignment to data sets, or cloud infrastructure to provide virtual server rooms (Dreyer, Döbler, & Rohde, 2015).
- **Repository services**, recommending external appropriate best practice discipline repositories or building own repository farms with specific focus to support high volume data formats found across many institutes at HU. It is essential to support interoperability with other

infrastructure and the natural workflows of scientists. One example is the development of a *media repository farm* covering all kind of media, like audio, video, graphics, texts which have to be grouped, enriched by metadata, integrated into the Moodle system for teaching or being cited in publications. Within the *Laudatio project* another flexible repository for historical linguistic corpora was developed and now is being used by more and more disciplines for corpora and annotation management (Krause et al., 2014). A *repository for tabular data* is currently in design and community building phase to support the management, visualization and alignment of tabular data.

- **Presentations and Workshops** for specific disciplines or stakeholder groups in the institutes designed for broad dissemination of the HU research data management activities. Workshops could be stakeholder focused, like for the Humboldt Graduate School doctoral students or more generic and introductory.
- A **Contact point** (person and web portal) has been established to aggregate all activities, information, tools and news ("DataMan - research data management portal at Humboldt-Universität zu Berlin," 2016).
- **Online tutorials** for data management will be created in 2016, like introductory movies, small modules to support the creation of research data management plans or metadata definition.
- **Developing strategic projects** to exploit further potential and applying for external funding in common projects between the infrastructural organizations of the HU and the institutes, like the DFG funded *eDiss+ project* to support doctoral students managing and publishing research data related to their thesis or the *re3data project* to list discipline specific repositories and their features worldwide.

The following figure provides a comprehensive overview about the components of the research data management concept:

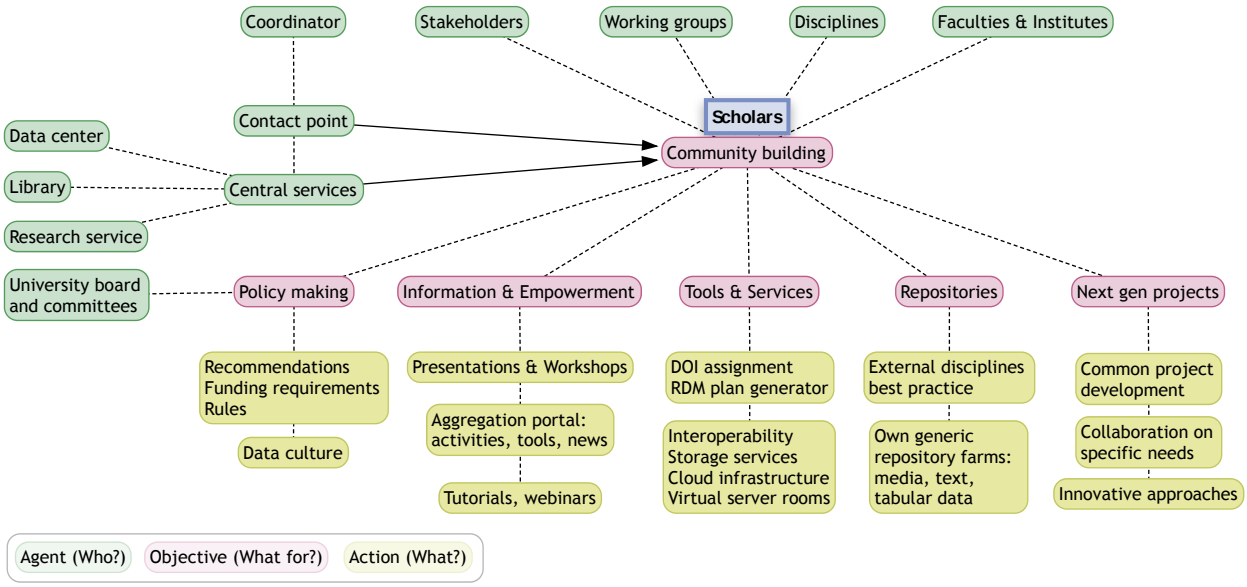

Figure 2: Agents, Objectives, and Actions

The position of the research data coordination and point of contact is situated within the data center's innovation group. Additional services are provided in the appropriate departments of the data center. The library advises on metadata enrichment and electronic publication aspects. The HU research

service center is providing legal advice about research data handling, e.g. about usage rights or intellectual property rights involved.

5. Community Building and Communication

As each scientific discipline has very unique communication, collaboration and publication cultures, the research data activities aim for a better understanding of their specific needs by stipulating discussions with the institutes and scholars most active in this area.

It is worth mentioning that each stakeholder group, professors, research fellows, project staff, or administration have a very distinct view on requirements and ways to address them. This makes multifold solutions to the very same problem necessary and raises the institutional awareness for research data, especially in the field of arts and humanities. To address these problems, discussions about specific faculty or institute positions for the local support of data management started. Currently, related support activities are very time consuming, like the development of small services, the cleaning of data, writing scripts to transform data or metadata structures. When it comes to shared staff for these tasks, the competitive requirements have to be managed in a very structured way to provide benefits for as many different people as possible and to keep the newly created position for research data coordination from congestion and finally dissatisfaction. Given the very specific issues within an institute's domain and therefore the long learning period involved, this new kind of position should provide an appropriate long-term perspective.

Additional communication activities include e.g. PhD seminars like "Introduction to research data management" or "How to develop a data management plan" as well as supporting scholars to find the suitable external research data repository.

We expect to fully establish this integrative approach by the end of 2017. New discussions about small supporting services and tools are currently started to include as much new insight from scholars as possible. As the cooperation with institutes intensifies and the community grows, the view on service gaps and further organizational instruments improves steadily. It is expected, that the current paradigm and trend about digitalization will amplify the needs for supporting services within the next 5 years. As research is ever changing and especially top research is changing focus every some years, it will be a big challenge to create instruments that are flexible enough and prepared for change as well as to stay close to state of the art.

6. Two examples for tools and services

To accomplish the formally adopted research data management policy and get its intentions into real-world action, it has to be accompanied by helpful services to enable for reliable, user-friendly management and ongoing curation of data. The university data center and the university library both continue to establish related services. The university library e.g. established services like the persistent identifier Datacite (Brase, 2009) service, inclusion into the library catalog or citation formatting as well as support and advice on metadata, digitization and additional identifiers. The data center currently focusses on repositories and storage technology. Both align and adjust their strategy in a common jour-fixe.

6.1. HU Box

One service created in the course of these activities is the "HU Box" service, which provides an online/offline sync and share scenario to HU users. At HU there are two main usage scenarios for such a service:

- a data exchange platform for network based sharing of files and collaboration of groups using various devices
- personal and synchronized data storage (file synchronization) with on-premise storage

Several free and commercial software products were evaluated by HU, JGU Mainz, and TU Kaiserslautern. The evaluation of the following software products was cancelled at HU due to missing

functions: aeroFS, Ajaxplorer, Druva inSync, Teamdrive and Tonido. Final candidates for closer examination were “Seafile”, “ownCloud” and “Powerfolder". Seafile turned out to be the most suitable solution for the requirements listed.

Besides the main scenarios for file exchange and sharing, additional more specifically research data oriented scenarios have been identified. These are e.g. in the fields of additional identifiers to be assigned as well as interfaces to existing systems, data privacy levels or collaborative editing features as well as write-once storage concepts to enable for stable data citation. They will be explored individually in separate projects.

6.2. Media Repository Farm

As another example service, the media repository farm will be explained in more detail. The media repository farm at the HU aims at providing a flexible, yet long-term supported media management platform for research and teaching. For this purpose, the open source software “ResourceSpace” was extended by modules enabling scholarly usage as well as teaching scenarios. A wide range of audio, video, textual or presentation formats is supported. The concept of “projects” is used in this farm for highly individual, self-administrable workspaces for HU users. The CMS of the HU provides a guarantee for at least 15 years of data and metadata longevity. The longevity of the underlying software platform “ResourceSpace” and related software development is dependent to many external factors. This is why a platform with a well-established and broad user base was selected instead of other very specialized software systems. The usage scenarios start from integrating contents in web content management systems or disseminating specific contents into the Moodle LMS as well as capturing photos while on the go with mobile devices and up to automatized imports of mass data with semi-automatic or even automatic extraction and enrichment of metadata. Currently e.g. an automated optical character recognition (OCR) is integrated in the media repository farm to better support the management of digitized media. As a well-established, still rapidly growing service, training for the media repository farm is included into the further education program of the HU on a regular base.

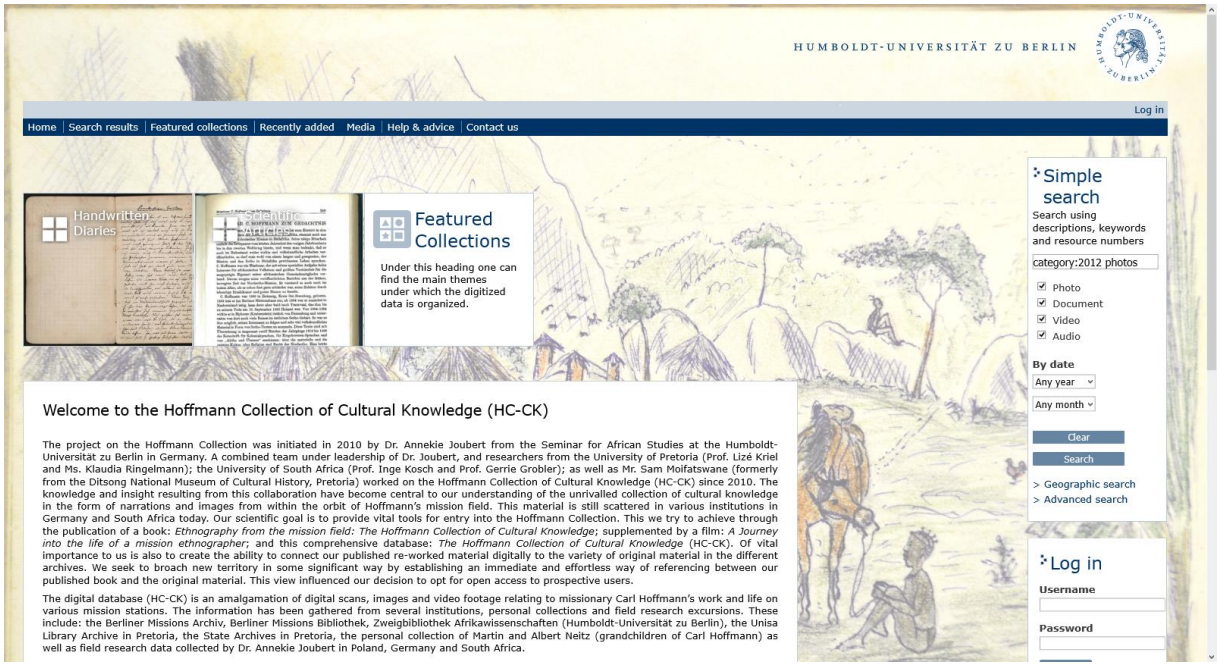

Figure 3: Media Repository Farm Example (Hoffmann Collection, http://dx.doi.org/10.17172/MR/22)

7. Conclusion

While research data management at the HU is still an evolving topic, a well-defined set of services to support scholars at the HU could be developed and put into operation. Having an established concept

for further improvements and developments for both, community building and services, helps to review the progress achieved so far. As there is no quick and comprehensive technological solution to research data management, communication and community work is still a key to increase the awareness and enhance related capabilities within the university. Especially for young researchers, the continuous workshop and training program is very well received.

Even with the mentioned set of community work, services and tools, the basic infrastructure to deal with research data in a transparent way, is far from being complete. Transparency in this context can be extended from pure data documentation to extended context descriptions and up to dynamic aspects of the creation, the software involved, activity protocols, related procurements and procedures. As these aspects will get workable, the requirements on research data services will constantly grow and will constantly demand more expertise from data centers to address them. It is expected that many, but not all related services will be provided by centralized, discipline specific service providers. Where not standardized (yet), local data centers will still be in charge.

For the HU data center CMS, the research data management activities have a strong impact on the service management procedures as the communication processes are closer to the scholars and often highly individual. So far, these changed communication styles had very beneficial side-effects by strengthening collaboration between the data center's service managers, administrators and the scholars. Within the coming years, support for research data management and closer involvement into the research data related processes also will have a strong influence on the service profile and innovation focus of the data center.

8. References

Arms, C., & Fleischhauer, C. (2005). Digital formats: Factors for sustainability, functionality, and quality. In *Archiving Conference* (Vol. 2005, pp. 222–227). Society for Imaging Science and Technology.

Berman, F., Wilkinson, R., & Wood, J. (2014). Building Global Infrastructure for Data Sharing and Exchange Through the Research Data Alliance. *D-Lib Magazine*, *20*(1/2). http://doi.org/10.1045/january2014-berman

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078.

Brase, J. (2009). Datacite-a global registration agency for research data. In *Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO'09. Fourth International Conference on* (pp. 257–261). IEEE.

DataMan – research data management portal at Humboldt-Universität zu Berlin. (2016). Retrieved February 29, 2016, from https://www.cms.hu-berlin.de/en/ueberblick-en/projekte-en/dataman-en/welcome?set_language=en

Dreyer, M., Döbler, J., & Rohde, D. (2015). Building Service Platforms using OpenStack and CEPH: A University Cloud at Humboldt University. *Eunis 2015*, *11*, 63–68.

Gross, A. G., Harmon, J. E., & Reidy, M. S. (2002). *Communicating Science: The Scientific Article from the 17th Century to the Present*. Oxford University Press. Retrieved February 29, 2016, from https://books.google.de/books?id=ctrhBwAAQBAJ

Hey, A. J., Tansley, S., Tolle, K. M., & others. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Microsoft research Redmond, WA.

Humboldt-Universität zu Berlin Research Data Management Policy. (2014, August 7). Retrieved February 29, 2016, from https://www.cms.hu-berlin.de/en/ueberblick-en/projekte-en/dataman-en/info/policy

Keralis, S. D., Stark, S., Halbert, M., & Moen, W. E. (2013). Research Data Management in Policy and Practice: The DataRes Project.

Krause, T., Lüdeling, A., Odebrecht, C., Romary, L., Schirmbacher, P., & Zielke, D. (2014). LAUDATIO-Repository: Accessing a heterogeneous field of linguistic corpora with the help of an open access repository. Digital Humanities 2014 Conference. Poster Session.

Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS. *International Journal on Digital Libraries*, *12*(4), 195–214.

Piwowar, H. (2013). Altmetrics: Value all research products. *Nature*, *493*(7431), 159–159. Retrieved February 29, 2016, from http://doi.org/10.1038/493159a

Schirmbacher, P., Kindling, M., & Simukovic, E. (2013). *Humboldt-Universität zu Berlin Research Data Management Survey Results*. ZENODO. Retrieved February 29, 2016, from http://dx.doi.org/10.5281/zenodo.7448

Schirmbacher, P., Kindling, M., & Simukovic, E. (2014). Unveiling research data stocks: A case of Humboldt-Universitaet zu Berlin. *iConference 2014 Proceedings*.

Schneider, R. (2013). Research Data Literacy. In S. Kurbanoğlu, E. Grassian, D. Mizrachi, R. Catts, & S. Špiranec (Eds.), *Worldwide Commonalities and Challenges in Information Literacy Research and Practice: European Conference on Information Literacy, ECIL 2013 Istanbul, Turkey, October 22-25, 2013 Revised Selected Papers* (pp. 134–140). Cham: Springer International Publishing. Retrieved February 29, 2016, from http://dx.doi.org/10.1007/978-3-319-03919-0_16

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., … Rennison, D. J. (2014). The Availability of Research Data Declines Rapidly with Article Age. *Current Biology*, *24*(1), 94–97. Retrieved February 29, 2016, from http://doi.org/10.1016/j.cub.2013.11.014

Warner, D., & Buschman, J. (2005). Studying the reader/researcher without the artifact: digital problems in the future history of books.

Wessels, B., Finn, R. L., Linde, P., Mazzetti, P., Nativi, S., Riley, S., and others. (2014). Issues in the development of open access to research data. *Prometheus*, *32*(1), 49–66.

9. Authors´ biography

Malte Dreyer is the technical director of the Computer and Media Service of Humboldt-Universität zu Berlin, Germany. Before, he was director for the department of research and development at Max Planck Society, Max Planck Digital Library.

He designed and developed research and publication data infrastructure for the Max Planck Society's institutes, as well as many research tools. Within several major German, European and international projects he is active in the areas of digital research infrastructure, repositories, virtual research environments and software architecture across many scientific disciplines.

Providing advice on software and information architecture, he is a member of several technical boards. He was a member of the German national alliance initiative working groups for research data and virtual research environments as well as a member of commission "Zukunft der Informationsinfrastruktur in Deutschland" and DINI member of the board.

Malte Dreyer's interests now are in the field of scalable information management architectures and infrastructures in the intersection of organizational perspectives on ICT from data centers and information management organizations. Current projects are in the fields of research data management and repositories, linguistics, as well as cloud architectures and integrated cloud services.

Andreas Vollmer has been teaching for a decade in the Humanities at Humboldt-Universität zu Berlin till he joined Computer and Media Service in 2002 when a new unit for the promotion of digital media in teaching and learning was formed. He coordinated this group and a number of projects focused on the development of services and infrastructural aspects.

He is interested in academic workflows and smooth interaction between its processes. Research data management has been assigned to his group to foster an integral implementation in existing infrastructure and new projects.