Status Report of bwFDM-Communities – A State Wide Research Data Management Initiative

Frank Tristram[*], Dennis Wehrle[**], Uğur Çayoğlu[*], Jessica Rex[***], and Dirk von Suchodoletz[**]

[*]Karlsruhe Institute of Technology, Steinbuch Centre for Computing (SCC)
[**]Albert-Ludwigs University Freiburg, Professorship in Communication Systems
[***]University of Constance, Communication-, Information,- Mediacenter

Abstract: Research data are valuable goods that are often only reproducible with significant effort or, in the case of unique observations, not at all. Scientists focus on data analysis and its results. By now, data exploration is accepted as a fourth scientific pillar (next to experiments, theory, and simulation). A main prerequisite for easy data exploration is successful data management. A holistic approach includes all phases of a data lifecycle: data generation, data analysis, data ingest, data preservation, data access, reusage and long term preservation. Tackling the challenge of increasing complexity in managing research data, the objective of bwFDM-Communities is to expose problems of research communities.[1] To achieve this goal, the project's key account managers enter into a dialogue with all relevant research groups at each university in Baden-Württemberg. Next to the identification of *best practices*, possible developments will be determined together with the scientists.

1 Project Motivation

Research data are generally understood as data that are generated during scientific work and they are building the basis for scientific results.[2] Such data can be very heterogeneous. They differ in origin, size and format, but share scientific significance. Research Data Management (RDM) includes all technical and organizational aspects of handling research data. This includes analysis, access, migration, integrity, metadata, visualization, and archiving, as well as cost models and legal aspects. In general, RDM attempts to increase scientific insight. This is usually difficult because research domains are very heterogeneous with respect to gaining knowledge and data demands. While some disciplines need to manage and analyze an enormous flow of data with each measurement they make, other disciplines only produce a few files in the course of their research. Therefore, most researchers do not focus on the management of their data. Nonetheless, the awareness that

[1]Especially the long tail of science, see e.g. `https://www.ci.uchicago.edu/blog/unwinding-long-tail-science` [last access 30.06.2014]

[2]DFG 2010 Call: "Informationsstrukturen für Forschungsdaten", `http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_forschungsdaten_1001.pdf` [last access 30.06.2014]

a lot of accessible research data is not consistent, poorly connected, not traceable, and can hardly be integrated into one's own work is something most researchers share.

The current European political vision concerning research data can be summarized by a statement from the European High Level Expert Group on Scientific Data for 2030: "Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure..." [3] A strong German and European Open-Access-Initiative is following this political incentive that availability of high quality research data is an important key for science in Germany and the whole of Europe. The European Union, Germany and the German Research Foundation have announced their research strategy and development plans in several publications.[3,4] These developments suggest that external funding bodies will increase their requirements for data management and support more "Open Data" projects in the future. This will be realized by adjusting funding standards to new demands as well as through gradually increasing the general demands for all publicly funded research projects. In the near future it is likely that an obligatory publication of primary data is requested from all such projects.[5]

The project bwFDM-Communities is directly getting into contact with scientific communities in order to estimate their demand for services and infrastructure to handle research data at Baden-Württemberg's universities. The objective of the project is to gain knowledge about how relevant research groups at Baden-Württemberg' universities handle their research data and to identify their problems in order to suggest best practices. No such qualitative analysis of similar size and depth has been attempted in Germany before.[6]

2 Project Structure

At each university in Baden-Württemberg a key account manager is established. The key account manager's main task is creating a link between infrastructure suppliers and IT service institutions, libraries, and other university research institutions. Furthermore, the key account manager tries to gather information about the demand of the scientists in a continuous dialogue. Thereby the full spectrum of scientific data from diverse research fields is covered. Engaging in a continuous dialogue with suppliers of infrastructure and

[3]Riding the wave – How Europe can gain from the rising tide of scientific data, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf [last access 30.06.2014]

[4]Concept for IT infrastructure in Germany, http://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Infrastruktur/KII_Gesamtkonzept.pdf [last access 30.06.2014]; Information processing in higher education – Organization, Services and Systems, Recommendations of the Commission for IT Infrastructure 2011-2015, http://www.dfg.de/download/pdf/foerderung/programme/wgi/empfehlungen_kfr_2011_2015.pdf [last access 30.06.2014]

[5]Horizon 2020 – Multi-beneficiary General Model Grant Agreement, http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf [last access 30.06.2014]

[6]Some universities have carried out single surveys among their scientists. The WissGrid project (http://www.wissgrid.de/) and the Nestor (http://www.langzeitarchivierung.de/) community previously tried to bring experts and scientists together to discuss RDM problems within the scientific communities in Germany.

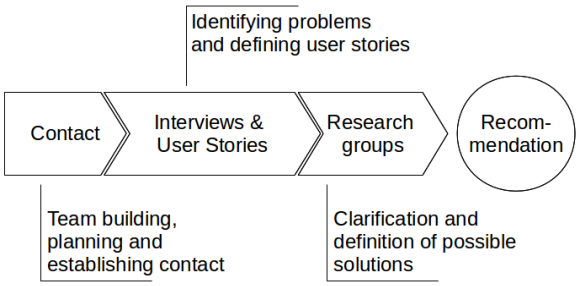

Figure 1: Phases of bwFDM-Communities project.

services the key account manager compiles all planned services. Ideally, the key account manager can hint at respectively helpful services and infrastructures for the scientist.

The project is divided into four phases (c.f. figure 1). The first phase, which has been reserved for team building, gaining knowledge, planning, establishing contacts and performing test interviews, ended in April 2014. The second phase, running from May 2014 till November 2014 is labelled as interview phase. Here, all nine key account managers are talking with scientists in order to identify scientific requirements and problems in connection with data management. Starting in December 2014, the clarification phase will follow, during which interested scientists and experts will come together in small groups and talk about possible solutions to specific problems that affect a larger number of scientists. In the last phase, which will cover the last three months of the project, a report will be drawn up and the results will be published.

Contact Phase In the first phase the key account managers' tasks were to identify the local data services at their universities, to create lists of all research groups at their respective universities and to identify a contact person for each research group. Furthermore, the next steps had to be planned and prepared taking special care of setting up an interview guideline properly. The guideline will be published at the end of the project.

Interview Phase Within six months, the current and future needs, problems and wishes of the scientists are identified. For this purpose a detailed conversation with each scientific group, concerning their present state and planned or expected developments, is necessary. As a result from the interviews user stories[7] will be written and – bringing together the various results – story maps[8] will be created.

Clarification Phase In the clarification phase topics of general importance as well as general demands will be defined and presented. This definition will be based on all user stories and the derived story maps as well as on the work done in the joint working groups.

[7] User stories are short paragraphs that explain a role, a wish (an instrument) and a purpose.

[8] A story map is a subsumption of related user stories that contains the main statements and also provides an embedding frame for each story.

Emerging questions and issues will be entered into a second questionnaire that mainly addresses technical details and aims at eliminating redundant stories efficiently.

Completion Phase The completion phase focuses on the evaluation, summary, and documentation of the project's results. On the basis of a substantiated demand analysis, detailed recommendations regarding concrete measures (e.g. building infrastructure, developing technology, transferring knowledge, ...), which can help to secure a sustainable research infrastructure for Baden-Württemberg's universities, will be provided. The recommended measures will feed into the second phase of the concept bwDATA as well as into the e-science strategy of Baden-Württemberg and complement the strategy by elaborating on concrete measures [HWC13].

3 Preliminary Findings

We have developed an interview guideline containing several elements we expected to come up frequently. Those can be marked like multiple choice answers by the interviewer. On the basis of these answers we will later build statistical plots and attempt to establish correlations. The structured interview covers the following topics of research data management: field of research, background and kind of data, data processing, storage, publication, data sharing and more general questions.[9] The qualitative information from the interviews is condensed into user stories. From a single interview typically two to three user stories, that include a specific need, requirement or problem, can be extracted by average. The user stories are stored in a database and reviewed by project members with regards to clarity and conciseness. The review is performed by joint working groups which are specialized on different topics and will ensure that the stories are well usable in the future.

As expected, our very first impression is that knowledge, structures, problems and solutions show a strong heterogeneity between the universities and disciplines. Some groups are not aware of the increasing importance and standards of data management set by the DFG or the EU. The experience from our interviews shows that even some of the professors do not know about DFG's guidelines and policies, a fact that has been proven by other studies before [SKS13]. Although data management plans are often formulated for project proposals, their implementation into the daily research routine is often missing. A reliable data storage system for a minimum of ten years is not provided in most cases. We also observed that some scientists do not care about what they write into their proposal. They write what they feel is expected of them but do neither take care of data management nor of the preservation of their data for at least ten years, as it is demanded by the DFG guidelines on good scientific practice. This may be partly caused by a lack of knowledge, motivation or incentive as well as a lack of reasons. This structural problem seems to be derived from unclear responsibilities. The scientist might see his institution in the position to care for

[9] E.g. whether the scientist feels well informed about research data management and which channel of information he or she would prefer.

a long term preservation of his or her data. The institution in turn, might not have the resources, or also might not feel responsible for providing long term access to all data. A solution might be university policies that clarify who is responsible for fulfilling the demands set by the funding agencies. Scientists are usually aware of all of these deficiencies themselves, for instance, the reputation of a small scientific group hardly increases when sharing their raw data, but they often worry about larger groups that might then dominate their field of research. Such dilemmas cannot be solved by official requirements alone. Apart from scientists lacking a thorough research data management, there are, of course, groups that show a high quality of data management.

However, aside from large collaborations, an exemplary data management behavior seems to either be directly important for the scientist's daily work or else depends on single persons who insist on good practices with research data. Proper tools, services and guidelines can at least create a very supporting environment for good practices and are also often requested by the scientists. Currently there is a perceived gap between demand and received support in research data management for small scientific communities.

Acknowledgments

The work presented in this publication is part of the project "bwFDM-Communities - Wissenschaftliches Datenmanagement an den Universitäten des Landes Baden-Württemberg" sponsored by the Ministry for Science, Research and Arts of the federal state of Baden-Württemberg, Germany.

The work would not have been possible without the support of many other people. The authors wish to express their gratitude to all additional key account managers: Dieta-Frauke Svoboda, Peter Bamberger, Jörg Hertzer, Johannes Knopp, Jonas Kratzke, Fabian Schwabe and Denis Shcherbakov. We furthermore want to thank Wiebke van Ekeris, Claudia Kramer, Pia Daniela Schmücker, Uli Hahn, Hans-Jürgen Goebelbecker and Karlheinz Pappenberger from the university libraries of Freiburg, KIT, Konstanz and Ulm for their time and contributions to our project. We are finally highly indebted to thank Marcus Hardt, Christopher Jung and Achim Streit from SCC at KIT for their supervision and for providing necessary information regarding the project. They have set up and guided the project in the lead time and the starting month.

References

[HWC13] Hannes Hartenstein, Thomas Walter, and Peter Castellaz. Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste. *PIK-Praxis der Informationsverarbeitung und Kommunikation*, 36(2):99–108, 2013.

[SKS13] Elena Simukovic, Maxi Kindling, and Peter Schirmbacher. Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. 2013.