Research Data Management in the Lab

Matthias Razum[*], Simon Einwächter[+], Rozita Fridman[*], Markus Herrmann[#], Michael Krüger[+], Norman Pohl[§], Frank Schwichtenberg[*], Klaus Zimmermann[+]

[*] FIZ Karlsruhe, Hermann-von-Helmholtz-Platz 1,
76344 Eggenstein-Leopoldshafen, Germany
{firstname.surname}@fiz-karlsruhe.de

[#] Stuttgart University Library, Holzgartenstraße 16,
70174 Stuttgart, Germany
{firstname.surname}@ub.uni-stuttgart.de

[+] Freiburg Materials Research Center (FMF), Stefan-Meier-Straße 21
79104 Freiburg, Germany
{firstname.surname}@fmf.uni-freiburg.de

[§] Stuttgart Media University, Nobelstraße 10,
70569 Stuttgart, Germany
{firstname.surname}@hdm-stuttgart.de

Introduction

Research, especially in science, is increasingly data-driven (Hey & Trefethen, 2003). The obvious type of research data is raw data produced by experiments (by means of sensors and other lab equipment). However, other types of data are highly relevant as well: calibration and configuration settings, analyzed and aggregated data, data generated by simulations. Today, nearly all of this data is born-digital. Based on the recommendations for "good scientific practice", researchers are required to keep their data for a long time. In Germany, DFG demands 8-10 years for published results (Deutsche Forschungsgemeinschaft, 1998). Ideally, data should not only be kept and made accessible upon request, but be published as well – either as part of the publication proper, or as references to data sets stored in dedicated data repositories. Another emerging trend are data publication journals, e.g. the Earth System Science Data Journal[1].

In contrast to these high-level requirements, many research institutes still lack a well-established and structured data management. Extremely data-intense disciplines like high-energy physics or climate research have built powerful grid infrastructures, which they provide to their respective communities. But for most „small sciences", such complex and highly specialized compute and storage infrastructures are missing and may not even be adequate. Consequently, the burden of setting up a data management infrastructure and of establishing and enforcing data curation policies lie with each institute or university. The ANDS project has shown that this approach is

[1] http://www.earth-system-science-data.net/

even preferable over a central (e.g., national or discipline-specific) data repository (The ANDS Technical Working Group, 2007). However, delegating the task of proper data curation to the head of a department or a working group adds a huge workload to their daily work. At the same time, they typically have little training and experience in data acquisition and cataloging. The library has expertise in cataloging and describing textual publications with metadata, but typically lacks the discipline-specific knowledge needed to assess the data objects in their semantic meaning and importance. Trying to link raw data with calibration and configuration data at the end of a project is challenging or impossible, even for dedicated 'data curators' and researchers themselves. Consequently, researchers focus on their (mostly textual) publications and have no established procedures on how to cope with data objects after the end of a project or a publication (Helly, Staudigel, & Koppers, 2003).

This dilemma can be resolved by acquiring and storing the data automatically at the earliest convenience, i.e. during the course of an experiment. Only at this point in time, all the contextual information is available, which can be used to generate additional metadata. Deploying a data infrastructure to store and maintain the data in a generic way helps to enforce organization-wide data curation policies. Here, repository systems like Fedora[2] (Lagoze, Payette, Shin, & Wilper, 2005) or eSciDoc[3] (Dreyer, Bulatovic, Tschida, & Razum, 2007) come into play. However, an organization-wide data management has only a limited added-value for the researcher in the lab. Thus, the data acquisition should take place in a non-invasive manner, so that it doesn't interfere with the established work processes of researchers and thus poses a minimal threshold to the scientist.

The BW-eLabs Project

Many disciplines require expensive and complex experimental resources, such as laboratory equipment. These resources are scarce and hard to obtain, especially for researchers at universities that do not own the equipment and for educational purposes. The German state of Baden-Württemberg has thus decided to fund the project BW-eLabs (Jeschke, et al., 2009), which is aiming at providing access to diverse remote and virtual experimental resources in combination with scientific data management. The field of nanotechnology is an adequate demonstrator discipline due to outstanding cost-intensive experimental equipment. (Remote) access to experimental equipment is an important prerequisite to the wide spread availability of professional tools for all scientists involved. Likewise the accessibility of gained knowledge, encompassing not only retrievability but also means for understanding, is fundamental. Advancement of cooperation and collaboration in scientific communities in high-tech fields takes centre stage in this notion.

Consequently, key concepts of the project include the reproducibility of experiments, discoverability of and access to primary data, and the storage and curation of all artifacts that emerge throughout the research process. The eSciDoc

[2] http://www.fedora-commons.org/
[3] https://www.escidoc.org/

Infrastructure (Razum,Schwichtenberg, Wagner, & Hoppe, 2009) provides such a data management system and forms one of the major building blocks of the BW-eLabs architecture.

Data Acquisition in the Lab

The nanoscience group[4] at the Freiburg Materials Research Centre (FMF) focuses on the synthesis and characterization of nanomaterials, their modification, and the implementation of (modified) nanomaterials into applications. One of the challenges is the reproducibility of the synthesis of nanoparticles. The particular difficulties in the classical chemists approach is the wide range of parameters. For example it is not uncommon to have specifically tailored glass equipment and strongly varying heating systems, which can only be controlled very roughly. Consequently, the group aims at automatizing the synthesis process by means of a remote-controllable microwave oven. At the same time, the "live" or online analysis of the synthesis process is pivotal to gain comparable measurements, which are not impaired by the inaccuracy of taking manual probes. Thus, two spectrometers (absorption and photoluminescence) are being integrated with the microwave oven to allow for contact-free, real time measurements. This allows for a continuous online analysis, resulting in dozens or even hundreds of spectra per experiment.

Both spectrometers store their measurements as files in the file system of a laptop used to control the equipment. The researcher creates a new experiment via the eLab Solution, which includes sending a configuration message to the eSync Daemon. The eSync Daemon then monitors the file system folder on the laptop to which the spectrometer writes the data files. The daemon replicates the files and sends them to a Deposit Service, which automatically extracts metadata from them, creates a new object by merging the data file with the metadata into an eSciDoc Item, and deposits the Item in eSciDoc as part of the experiment folder. Also a QR-Code printout is generated by the eLab Solution in order to refer to the experiment data from a laboratory journal. The data acquisition workflow and its components are depicted in figure 1.

The whole system is Java-based (except for the user interface of the eLab Solution, which is based on a JavaScript framework). It is independent of the underlying operating system. The Metadata Extractor is based on Harvard's FITS service[5] (Stern & McEwen, 2009) and can be extended to work with other file formats by providing appropriate plug-ins. The Deposit Service currently relies on eSciDoc as a data sink, but may be adapted to work with other repositories as well. The communication between the data acquisition components is fully REST-based. This simplifies the extension of the system and the integration of other lab equipment.

In enhancement of the metadata extraction the complete available information shall be retained, besides the unmodified original data in whichever format it will be presented by the lab equipment, in the Full Metadata Format (Riede, Schueppel, Sylvester-Hvid, Kühne, Röttger, Zimmermann & Liehr, 2009). This guarantees not

[4] http://www.fmf.uni-freiburg.de/projects/pg_anorg_en/AKKrueger
[5] File Information Tool Set, see http://code.google.com/p/fits/

only the long term reusability of the acquired knowledge, but also, due to its simple format, the applicability of a wide range of tools.

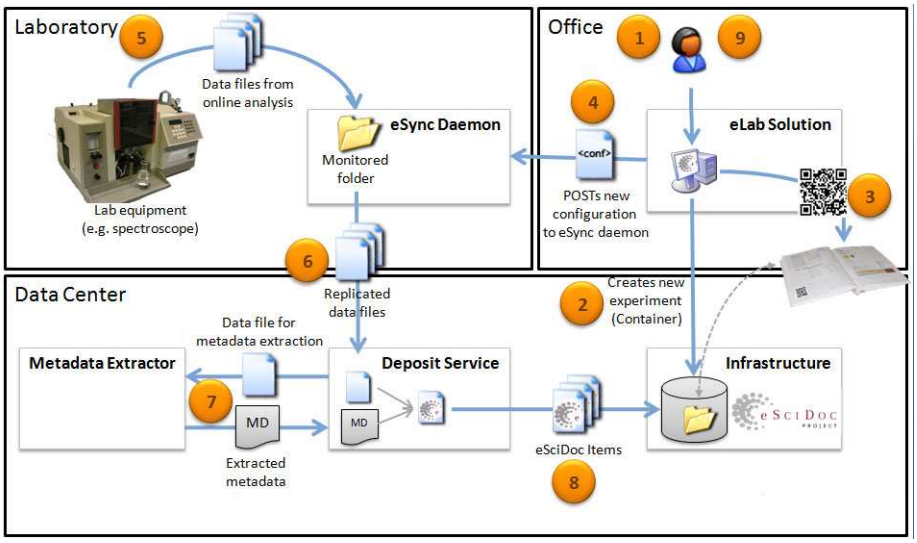

Figure 1: The data acquisition workflow in the nanomaterials lab with the underlying eSciDoc data infrastructure and the eLab Solution to configure experiments and access the gathered data objects.

Future Work

The work presented in this abstract is only a first step towards a fully automated data acquisition workflow for the nanomaterials lab. The next step will be a deeper integration of the lab equipment with the underlying data management software and the BW-eLabs portal to interface directly the lab equipment. This will not only allow accessing additional calibration and configuration data (especially for the microwave oven), but enable researchers to configure the microwave oven (and later the spectrometers) from within the eLab Solution and the BW-eLabs portal. This will permit to run complex experiments remotely with minimal support from the local lab staff.

Further work will generalize the concepts and applications to facilitate the roll-out of the eLab Solution for other working groups within the Freiburg Materials Research Center and the re-use in other disciplines and research institutes.

References

Deutsche Forschungsgemeinschaft. (1998). *Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“ (Denkschrift).* Weinheim: Wiley-VCH.

Dreyer, M., Bulatovic, N., Tschida, U., & Razum, M. (2007). eSciDoc - a Scholarly Information and Communication Platform for the Max Planck Society. *German e-Science Conference.* Baden-Baden.

Helly, J., Staudigel, H., & Koppers, A. (25. 01 2003). Scalable models of data sharing in Earth sciences. *Geochem. Geophys. Geosyst. , 4* (1), p. 1001.

Hey, T., & Trefethen, A. (2003, May 30). The Data Deluge: An e-Science Perspective. *Grid Computing* , pp. 809-824.

Jeschke, S., Burr, B., Hahn, J. U., Helmes, L., Kriha, W., Krüger, M., et al. (2009). Networking Resources for Research and Scientific Education in BW-eLabs. In *10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing* (S. 47-52). IEEE Computer Society.

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2005, December 29). Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries , 6* (2), pp. 124-138.

Razum, M., Schwichtenberg, F., Wagner, S., & Hoppe, M. (2009). eSciDoc Infrastructure: A Fedora-Based e-Research Framework. In M. Agosti et al., *ECDL 2009, LNCS 5714* (p. 227-238). Springer.

Stern, R., & McEwen, S. (2009). FITS – The File Information Tool Set. *OR09. Conference Posters* (http://hdl.handle.net/1853/28508). Atlanta: Georgia Institute of Technology.

Moritz Riede, Rico Schueppel, Kristian O. Sylvester-Hvid, Martin Kühne, Michael C. Röttger, Klaus Zimmermann & Andreas W. Liehr (2009). On the communication of scientific data: The Full-Metadata Format (http://dx.doi.org/10.1016/j.cpc.2009.11.014). In *Computer Physics Communications*, Vol. 181, Issue 3, March 2010, pp. 651-662.

The ANDS Technical Working Group. (2007). *A proposal for an Australian National Data Service.* Canberra: Australian Government, Department of Education, Science, and Training.