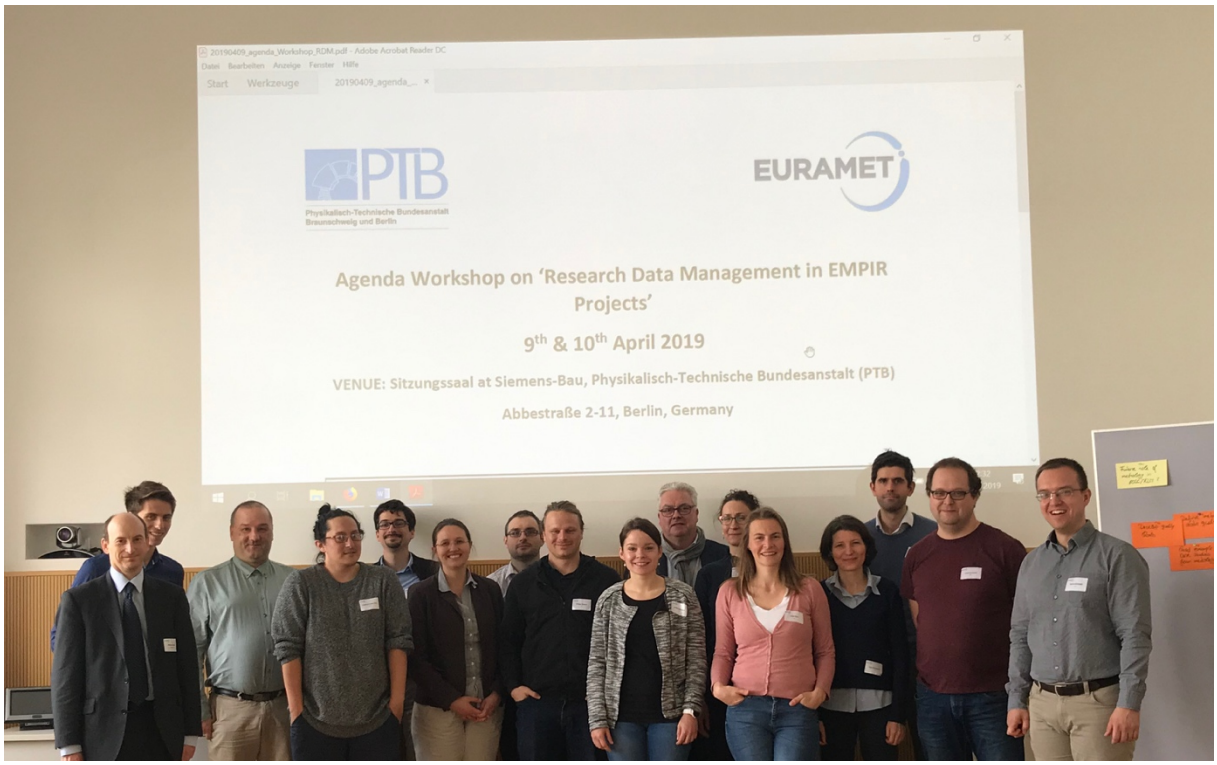
Workshop on ‘Research Data Management in EMPIR Projects’

9th & 10th April 2019

VENUE: Sitzungssaal at Siemens-Bau, Physikalisch-Technische Bundesanstalt (PTB)

Abbestraße 2-11, Berlin, Germany



Agenda

Day 1

1. Introduction (Sascha Eichstädt, PTB)
2. Metadata and information retrieval basics (Julia Neumann, PTB)
3. Technologies and platforms for RDM (Giacomo Lanza, PTB)
4. Presentation from running EMPIR projects
5. Discussions on metadata, RDM platforms and DMP (All)

Day 2

6. What EURAMET expects from EMPIR projects w.r.t. RDM (William Dawson, NPL)
7. Discussion of RDM strategies, guidance and technology

Workshop aims and scope

This workshop brought together several of the EMPIR projects that already have to develop and maintain a data management plan (DMP) in accordance with EU Horizon2020 rules. In the workshop participants presented their approach to research data management, publication, data repositories and metadata. The aim was to exchange experiences, strategies and information about useful technologies.

The workshop organization was aligned with the aims of TC-IM 1449 to foster the development of harmonised research data management (RDM) and metadata standards for metrological data and services. This is the requirement for establishing a joint metrological implementation network of the European Open Science Cloud (EOSC) principles of FAIR data and services. (Findable, Accessible, Interoperable, Reusable).

Information on areas around research data

The workshop started with two talks on basic knowledge related to data management. These talks were given by Julia Neumann (metadata and information retrieval) and Giacomo Lanza (technologies and platforms) from PTB. The following sections provide a brief summary of the contents of these talks.

Metadata and information retrieval

The computer-based extraction of information is called Information Retrieval (IR). One of its main purposes is to find relevant information to the users. Depending on the information seeking behaviour, and the over-all context of the end users, many different approaches can be applied and combined to offer a good information seeking experience. Most of the IR-methods can be set up within running digital information infrastructures without interrupting existing and established working processes too much. Search engines such as Google, DuckDuckGo, Bing etc. are popular established supplies and examples of Information Retrieval.

Before data can be extracted it needs to be processed. AI and machine intelligence approaches already help automatizing the processing of information. However, they cannot support an entire Information Retrieval workflow yet. They are neither flexible enough to deal with the many human emotion driven information needs, nor are they easy to handle because a big training sample is required. This isn't usually the case for most of the projects. Therefore, most of the Information Retrieval workflows are based on half-automatized methods such as the development of thesauri for controlled vocabulary, quantitative/qualitative interviews for the detection of user needs and metadata schemes to describe the information formally in a machine understandable manner.

The most important factor that needs to be considered is how to keep the balance between order, flexibility and automatization in general. One of the many discussion topics during the workshop focused on the question how to engage users in the process of data processing and how to decrease the workload. Discussed approaches involved the availability of well phrased drafts that may serve as an example which explains how to fill out e.g. metadata fields. Other approaches involved the need for data processing that centers rather on the motivation behind information seeking behaviour than the mere division between formal and content driven data processing behind IR. All in all, the data should be prepared beforehand in the best way, so that regular users who normally don't participate in data management get a satisfactory guidance. Describing them and thinking about these aspects

during the development of a RDM plan should happen at a very early stage as a precondition for the defined data management goals.

Technologies and platforms

The talk gave an overview of the practical aspects of research data management, namely the arising tasks and some solutions to perform them in an effective way.

In the introductory section, starting from the new challenges arising from today's data-driven science, the well-known FAIR principles were analysed, yielding concrete requirements to ensure machine-readability of research data. The aspects considered were: file formats; dataset description with metadata; usage licences; resource referencing with persistent identifiers.

In the following some tasks were discussed in detail, together with some tools which can come at help:

- Documentation of workflows by means of electronic lab books (e.g. Labfolder).
- Metadata enrichment of datasets with annotation software (RightField).
- Short-term storage of working documents, including data protection, cooperative working and quality control, by means of databases or software for versioning / continuous integration (e.g. GitLab).
- Long-term storage of definitive data, including data securing, sharing and possibly access control; the features of different data repositories and arguments for and against publications were presented.

The last part of the talk was dedicated to the data management plan, a document which helps a project leader doing research data management in a conscious and organised way. A valuable help for that purpose are tools such as the RDMO (Research Data Management Organiser), which allows to input all project- and data-related information in a structured way via a user-friendly questionnaire and to export it as a document according to the funder's layout, or to retrieve it through a database query. Therefore, PTB has implemented in the RDMO instance hosted at PTB an export scheme especially for EMPIR projects. During the workshop it was agreed with MSU and EMPIR programme management to jointly work towards an output from the RDMO that is acceptable by the MSU without any further adjustment.

As a conclusion, it was illustrated that research data management does not necessarily mean more work to do, if the right tools are provided. In fact, during the discussions at the workshop it was recognised that the writing of a DMP helps to think through the project's handling of research data in advance. In this process it is important to remember that the DMP to be submitted to MSU only needs to consider the data being made publicly available.

Methodology demonstrator

The utility of rich metadata description was also demonstrated with a pilot application, mimicking a small data repository with 10 artificial numerical datasets. The description of the datasets included, besides common metadata such as author and title, some information concerning the contained values: quantity name (taken out of a controlled vocabulary), unit (multiple choice dependent on the quantity), minimum and maximum value. Addressing these additional fields with an advanced search function allows answering complex research questions (e.g. "find all experiments with reported temperature values between 0 K and 3000 K") selecting only relevant datasets and dismissing false positives.

Physikalisch-Technische Bundesanstalt
Braunschweig und Berlin

Best-practices from projects from EMPIR Call 2017

In an open round of presentations, several attendees presented data management in the EMPIR projects they are involved in. From these presentations and the corresponding discussions, the following best-practices have been identified. These will soon be summarised into a document to support EMPIR projects starting this year and beyond.

Define what is „data“ for your project

In order to streamline the DMP development process it is advisable to start with a common understanding of what kind of data you are considering in your project. Does it contain software, questionnaires, drawings or simply plain text based files of numerical data? Do you only consider data to be published accompanied with a scientific paper or do you also provide data sets as individual publications?

Consider the DMP as help for you

Filling out the entries in the DMP are in fact a good starting point to think about your measurements, the intended audience and to identify potential weak spots in the project plan. It thus advisable to not consider the DMP as an unnecessary administrative document, but as helpful guideline for starting a discussion with the project consortium. In addition, it is important to remember that the DMP considers only the data that will be made publicly available. Nevertheless it can be of great help for the successful and frictionless collaboration within the project to discuss data handling before its publication, too.

Think about metadata as early as possible

Metadata is data about data and thus the most important aspect when somebody wants to find your valuable data sets. It is advisable to think about potential hierarchies of metadata in your project. This can either mean structuring the definition of common metadata corresponding to its importance, i.e. necessary and supplementary. The other option is a hierarchy w.r.t. position in the data structure itself, i.e. metadata for a single measured value vs. metadata for the whole data set. Such a hierarchy helps to focus the DMP part about metadata by considering in particular the elements at the top of the hierarchy. For these elements it is also advisable to consider setting up an agreed vocabulary using standardised schemes where available.

Generate metadata as early as possible

Important information can get lost when generating metadata just before submitting the data set to a public repository. It is thus advisable to generate reasonable metadata as early as possible in the data lifecycle. This can be achieved most efficiently by thinking about the later use of the data set as early as possible. Again, the DMP can help with that from the beginning.

Good research data management starts in the lab

Publication of data sets valuable to others requires a well thought through data generation process. This starts in the lab where the data is measured and, ideally, the first metadata is created. It goes on with measures to guarantee data integrity and traceability. For later use of the data by others it also advisable not to leave important information about the data in the comments within the data analysis code written by the data experts. In addition, already in the lab data should be converted or acquired in open text-readable formats. This greatly improves later use of the data and is an important aspect of the FAIR principles.

Think about the location of your data early

Where should data from the project be published? Answering this question should be started by thinking about the intended users of the data. It is often preferable to publish in specific repositories than a generic one. The reason for this is that a specific repository usually offers richer metadata and thus, makes it easier to find the data.

Keep all partners motivated to think about data

The data handling shouldn't be left to a single person, but should be an integral part of the project organisation and implementation. It is thus advisable to set up a data access committee which overlooks the data before publishing. At the same time it is important to keep things as simple as possible for the partners. This could begin with training on the meaning of data management, the FAIR principles and the DMP early on in the project. In addition, a checklist for partners generating data sets can help to streamline data processes early in the data lifecycle.

Potential tasks and questions for TC-IM 1449

During the discussions at the workshop several issues came up which will be fed into the workplace of the project TC-IM 1449, which will have its strategic meeting mid June 2019.

The inputs from the workshop are as follows.

- Development of a „getting started" guide for EMPIR projects.
- Best practices in using a data repository with Zenodo as example
- Definition of minimal requirements for data repositories for EMPIR projects
- Definition of potential types of data to be considered in EMPIR projects

In addition, the workshop identified various points that shall be addressed by TC-IM 1449 in order to set the ground for establishing metrology as an anchor of trust in research data management.

- ❖ Development of quality labels for research data based on metrics for data quality
- ❖ Publication of good example case studies from metrology
- ❖ Establishing reproducibility in research through traceability in the research process
- ❖ Outline potential infrastructures for implementation and realisation of good practice in data management

Furthermore it is necessary to develop a joint vision for research data management in metrology and the role of metrology in the European Open Science Cloud (EOSC) infrastructure. This process may be initiated by identifying the benefits from the EOSC for metrology.

Further reading and material

- List of community metadata standards: http://www.dcc.ac.uk/resources/metadata-standards or http://rd-alliance.github.io/metadata-directory/
- Tidy data (among others, with values in columns): https://www.jstatsoft.org/article/view/v059i10/
- CSVY format (CSV with YAML header): http://csvy.org/ or https://blog.datacite.org/thinking-about-csv/