1st Conference on Research Data Infrastructure

Enabling RDM

https://doi.org/10.52825/CoRDI.v1i.289

Published: 07 Sept. 2023

Research Data Publication at Large Scale

Thomas Zastrow[1] [https://orcid.org/0000-0002-2844-1495] and Nicolas Fabas[1] [https://orcid.org/0000-0002-2224-0780]

[1] Max Planck Computing and Data Facility

Introduction

The MPCDF is the high performance computing center of the Max Planck Society. Beside hosting compute clusters, around 300 petabytes of research data are stored at the MPCDF. Many of these datasets have a size of several terabytes up to petabytes and are stored over a heterogeneous landscape of storage systems. The datasets are covering a wide variety of scientific disciplines, many different data formats and access restrictions. For these reasons, it is not possible to offer one centralized data publishing solution for the datasets. Instead, the MPCDF developed a flexible data publishing concept (see Fig. 1), taking into account the challenges mentioned above. In this paper, we will describe the two parts of our data publishing concept: first, the creation and handling of metadata and second, the operating model for data repositories.

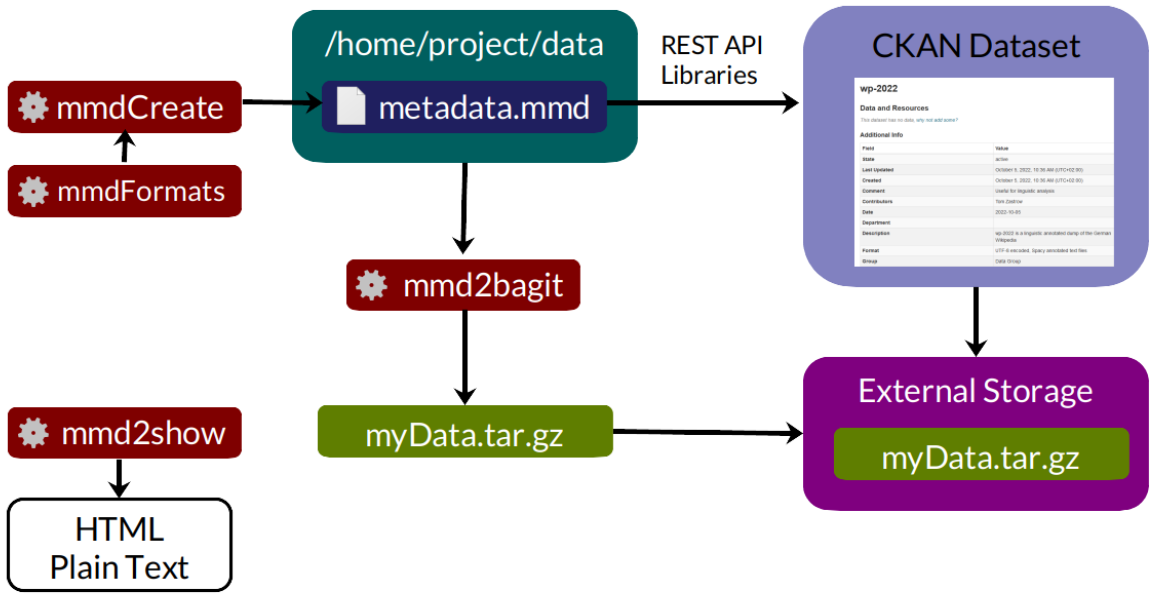

Figure 1: Research Data Publishing Workflow

Metadata Handling: The Metadata Tool Suite

If research data should be published in a data repository, descriptive metadata is necessary. This includes both administrative metadata (Who owns the data? Where is it stored?) as well as content related metadata (What is the data about? Who is allowed to access the data?).

Some of this metadata can be gathered automatically, but some information needs to be added manually. To support the latter type of metadata, the MPCDF developed the “MPCDF Metadata Tool Suite” (MMD Tools)[1]. The tool suite contains command line scripts which are flexible to be adapted and extended by the researchers themselves. The created metadata is stored in standardized JSON files, which allows a transformation into other formats as well as importing it into a data repository. Several common metadata formats are supported out of the box: currently, these are Dublin Core[2], the DataCite Metadata Schema (version 4.4)[3] and our own custom format which covers the basic needs of metdata handling at the MPCDF. But the MMD Tools are not restricted to these formats: supported by the MPCDF, researchers can create their own metadata schemata.

In detail, the tool suite contains the following scripts:

- **mmdCreate**: can be used to interactively create a new metadata file or edit an existing one
- **mmdShow**: displays or converts the content of a metadata file
- **mmd2bagit**: creates a self describing BagIt[4] container out of a given metadata file and the corresponding object data

In addition to the executable scripts, the tool suite contains a Python module with common functionality to be included in individual workflows. Below, a simple example of the stored metadata in Dublin Core format is shown:


```
{
    "id": "c8f6ea3d-c9ce-4cc4-99e1-7d36a7ed014f",
    "mmdFormat": "Dublin Core",
    "Format": "UTF-8 Encoded Text Files",
    "Type": "Annotated Text Corpus",
    "Language": "DE",
    "Title": "The wp-2022 Corpus",
    "Subject": "annotated text corpus,POS tagging,lemmatization",
    "Coverage": "Dump of the German Wikipedia from January 2022",
    "Description": "The corpus contains token, pos and lemma annotations",
    "Creator": "Dr. Thomas Zastrow",
    "Publisher": "---",
    "Contributor": "---",
    "Rights": "Creative Commons Attribution-ShareAlike 3.0 Unported License
(CC BY-SA) and the GNU Free Documentation License (GFDL)",
    "Source": "https://wikipedia.de",
    "Relation": "---",
    "Date": "2022-01-01"
}
```


[1] The MMD Tools are freely available under Apache 2.0 license and can be downloaded from the MPCDF GitLab. Installation instructions and further documentation can be found here: https://docs.mpcdf.mpg.de/doc/data/publication/mmd.html

[2] https://www.dublincore.org/specifications/dublin-core/

[3] https://schema.datacite.org/

[4] https://www.rfc-editor.org/info/rfc8493

Data Repositories: From Generic to Specific

Data repositories are a common way to publish research data in a FAIR way. At the MPCDF, the heterogeneous kind of the data stored makes it difficult to provide one repository for all the datasets. Therefore, the MPCDF developed a two-step procedure to fulfill the needs of the Max Planck institutes. The basis is a repository software off the shelf, which is installed and initially configured by the MPCDF. After that, the repository is handed over to the institute. Any adaptions and further configurations can then be done by the researchers or IT staff of the responsible institute. The MPCDF is still supporting the ongoing work with practical help in any directions, but, after initialization and configuration, the responsibility lies at the institute.

As there are many (open source) software solutions available which can be used as basis for a data repository, the MPCDF decided to go with the “Comprehensive Knowledge Archive Network” (CKAN)[5]. For our purposes, the CKAN software has several advantages. One key feature is the possibility to store only metadata in the repository and leave the object data where it is: datasets at the MPCDF are often too big to be easily moved around or to get downloaded via a web browser. Another feature is the extensive REST-style API which can be used for automation from nearly any programming language. Built on top of the REST API, wrapper libraries for several programming languages are available. In addition to the basic functionality of the CKAN software, the MPCDF supports the use of some extra plugins and developed specifically a plugin to assign DOIs to a dataset using MPG’s DOI service.

Any CKAN instance can store metadata following one or more schemata. These metadata schemata can be harmonized with the metadata created by the MMD Tool suite as mentioned above. By utilizing the repository’s API or implementing the available programming libraries, automated metadata and data publishing workflows are possible.

Currently, the MPCDF hosts several individual data repositories for Max Planck institutes. They are covering a range of scientific disciplines from archeology, physics (dark matter research and physics of light) to material research (polymer research). In some use cases, the data repository is part of a bigger (automated) workflow. This includes for example the integration of Jupyter Notebooks as well as the integration of HPC systems for analysing the data.

Conclusion

The combination of the MMD tool suite for metadata handling and CKAN as repository software is an easy and flexible solution for publishing heterogeneous research data. It allows the researchers to quickly and conveniently add metadata to their project’s data in the same environment they are using for research. After that, an adapted and highly configurable data repository allows the publication of research data in a FAIR way.

[5] https://ckan.org/