OPINION ARTICLE

Related to other papers in this special issue	23 (p230); 24 (p238); 25 (p246); 27 (p264)
Addressing FAIR principles	F, A, I, R

FAIR Practices in Europe

Peter Wittenburg[1†], Michael Lautenschlager[2], Hannes Thiemann[2], Carsten Baldauf[3] & Paul Trilsbeek[4]

[1]Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany

[2]DKRZ Ringgold standard institution, Bundesstr. 45a, Hamburg, Hamburg 20146, Germany

[3]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, Berlin 14195, Germany

[4]MPI for Psycholinguistics, Wundtlaan 1, Nijmegen 6525 XD, The Netherlands

Keywords: Infrastructure; FAIR Metrics; GO FAIR Matrix

Citation: P. Wittenburg, M. Lautenschlager, H. Thiemann, C. Baldauf & P. Trilsbeek. FAIR practices in Europe. Data Intelligence 2(2020), 257–263. doi: 10.1162/dint_a_00048

ABSTRACT

Institutions driving fundamental research at the cutting edge such as for example from the Max Planck Society (MPS) took steps to optimize data management and stewardship to be able to address new scientific questions. In this paper we selected three institutes from the MPS from the areas of humanities, environmental sciences and natural sciences as examples to indicate the efforts to integrate large amounts of data from collaborators worldwide to create a data space that is ready to be exploited to get new insights based on data intensive science methods. For this integration the typical challenges of fragmentation, bad quality and also social differences had to be overcome. In all three cases, well-managed repositories that are driven by the scientific needs and harmonization principles that have been agreed upon in the community were the core pillars. It is not surprising that these principles are very much aligned with what have now become the FAIR principles. The FAIR principles confirm the correctness of earlier decisions and their clear formulation identified the gaps which the projects need to address.

† Corresponding author: Peter Wittenburg (E-mail: peter.wittenburg@mpcdf.mpg.de, ORCID: 0000-0003-3538-0106).

1. INTRODUCTION

The FAIR principles [1] have now been widely accepted as guidelines of how to create, manage and curate data and services across research disciplines and organisations. Large infrastructure initiatives such as EU Open Science Cloud [2] state clearly that they expect contributions to this integrated domain of data and services to be FAIR compliant. In major research organisations in Europe such as the Max Planck Society, the awareness is growing that following the FAIR principles will facilitate data-intensive science (DIS), in which frequently data is used that comes from many different sources. The FAIR principles make explicit what has been understood already in various data intensive science communities for many years, but they are more consistent in the way they formulate the requirements for modern DIS than earlier attempts. Three examples from different Institutes within the Max Planck Society focussing on humanities, environmental and materials science are used to indicate how large research organisations are trying to meet the challenges of DIS and to meet the requirements of FAIR data as efficiently as possible.

2. HUMANITIES SCIENCE: DOBES PROJECT AND ARCHIVE

In 2000, the DOBES[1] project on Documenting Endangered Languages (http://dobes.mpi.nl) was started, finally including 75 multidisciplinary teams with on average 3 researchers from all over the globe with different backgrounds (from linguists to ship builders) managed to document about 120 endangered languages following agreed documentation guidelines. In the beginning the partners agreed to set up a central repository that will receive one copy of each created digital object, in order to preserve the material and to share it with the interested global community including the communities speaking the languages themselves. This repository is now a part of The Language Archive at the Max Planck Institute for Psycholinguistics, which is certified according to the Core Trust Seal [3] requirements for Trustworthy Data Repositories. Since everyone in the project understood the value and unique character of the cultural heritage that was being collected, much effort was taken by the archiving and documentation teams to carry out all work according to a set of guidelines that would make data FAIR[2] and accessible over long periods of time.

Early on in the project, a rich metadata schema (IMDI) [4] was developed together with the interested community, a metadata editor was developed supporting the schema, a repository structure was designed to support efficient data management and a PID server was set up. This facilitated the implementation of a repository system that supported FAIRness. At data upload time IMDI compliant metadata was created or uploaded, a globally resolvable PID was registered and added to the metadata and 4 external copies of all uploaded digital objects were created dynamically. In addition, all metadata was made harvestable through OAI-PMH, HTML versions were created to support web search engines, a browsing and search tool was developed, and tools were built to enable annotations of time series (media streams, EEG, eye tracking, etc.) and their visualisation according to open standards. All metadata was open and also data was made

[1] **Do**kumentation **Be**drohter **S**prachen.

[2] At that time the term "FAIR" was not known.

open when possible, however, due to the high sensitivity of some of the data, 3 additional access levels have been defined. All data users need to accept a Code of Conduct [5], since no common legal basis could be defined for working with this data.

In addition to being widely compliant with the FAIR principles including the definition of semantic categories being used in metadata and in linguistic description, much effort was put in automatic quality assessment and correction, in preservation aspects and in ethical correctness, i.e., about 10 repositories were established in various countries around the globe to store copies of language materials recorded in the corresponding regions.

Due to its careful systems design and the various tools supporting the users, the DOBES Archive is a great resource for researchers globally who are interested in the diversity of languages and cultures. Data of different types were integrated based on clear standards, and community best practices were defined, which were re-used in other projects worldwide. By overcoming the fragmentation in the field, researchers could now start addressing new types of scientific questions, for example about the evolution of languages, about different strategies hidden in languages to achieve the required expressiveness, about differences in intonation contours, about different phoneme systems, etc. With the availability of online repositories, these analyses can now be based on data and not only on publications by other researchers.

3. ENVIRONMENTAL SCIENCE: WORLD DATA CLIMATE CENTER (WDCC)

The WDCC hosted by the German Climate Computing Center (DKRZ) (www.wdc-climate.de) is part of the global climate modelling community that needs to support the delivery of the IPCC Assessment Report No. 6 [6] and the scientific goals of the global earth-system modelling community. To support this work the CMIP (Climate Modelling Intercomparison Project) data standard is being maintained and currently being extended to version 6. While CMIP5 includes about 2 PB uncompressed data spread over 5 Mio files, CMIP6 [7] will include 20 PB of compressed data stored in 50 Mio files. WDCC's mission is to carry out proper data management, i.e., long-term archiving, cataloguing, curation, and publishing of climate model output from globally distributed providers. The WDCC is certified according to the **Core Trust Seal** [8] a joint standard developed in RDA.

After having applied extensive quality assessment, **Findability** of WDCC data is ensured by publishing them via DataCite [9] which implies an assignment of DOIs at coarse granularity. These aggregations include hierarchical collections with metadata describing all levels in that hierarchy and Handles assigned to all individual data objects. Rich metadata is created following community standards as part of CMIP and is stored in a relational database. All metadata records are available for external harvesters through an OAI PMH interface and a mapping to the Dublin Core [10], ISO 19135 [11] and DataCite [12] XML metadata standards is provided. Various harvesters such as EUDAT B2FIND [13] are aggregating these metadata and the local WDCC GUI [14] also offers search and browse capabilities.

Accessibility of WDCC metadata is given by supporting OAI-PMH and data can be retrieved by HTTP which are both open, free and universally implementable standards. While metadata is openly accessible, access to data requires a registration at WDCC. Authentication is being checked using well-established methods. Terms of use [15] are provided. When data is being lost or removed, the metadata will indicate the reason.

To guarantee **Interoperability** of metadata, internal rich metadata is being mapped to well-defined standards such as Dublin Core, ISO 19135 and DataCite XML schema. Data are using open format standards [16] and also the data model of the CERA database is documented publicly [17]. Most of the climate data conform to the CF-netCDF standard which is in general self-describing and machine-readable and relies on commonly used controlled vocabularies. Other formats need to be supported as well to not discourage users from depositing valuable scientific data, but carrying out formal checks and other services are not applied. CF–netcdf structure and vocabularies are publicly documented and openly accessible. There is the intention of the CF committee of making the conventions citable via DOIs. Relations that can be specified using the DataCite "relationType" attribute are implemented in CERA and are accessible from the CERA web user interface and the harvesting interfaces. Users are supported in providing relations as relevant and possible for their data.

To improve **Reusability**, WDCC metadata, in general, contain rich information about the context in which data was generated such as relevant timestamps (creation and collection date), conditions under which data were created, actors involved in preparing the data, and model-related technical attributes such as model parameters and model descriptions. Due to complexity reasons not all lend for machine interpretability, i.e., accuracy descriptions may be provided as free text. Metadata is released under CC0 universal license terms. The data licenses are dependent on the data provider. However, WDCC recommends using CC-by 4.0. Quite a number of essential provenance attributes are stored in the CERA metadata. However, provenance information related to the workflow or procedures involved in generating data is only described at a basic level with project and experiment summaries and accuracy reports.

Summarising we can state that the workflows and principles being applied by WDCC are generally modelled along FAIR and have later benefitted from the principles. The priority to ingest all kinds of data types even if they are not compliant to widely accepted standards guarantees to capture important climate data that may be relevant in future.

4. MATERIALS SCIENCE: NOMAD REPOSITORY AND ARCHIVE

The Novel Materials Discovery (NOMAD) Laboratory (https://nomad-coe.eu) handles the worldwide largest repository of materials science simulation data. To that end, the NOMAD Repository accepts uploads of input and output files of the relevant computer programs of this community, processes the raw data and makes data available also under consideration of analysis by means of artificial intelligence methods [18].

In 2013, the NOMAD Repository was jointly started by Humboldt-Universität zu Berlin and the Fritz-Haber-Institut der Max-Planck-Gesellschaft in Berlin. This endeavour led, in fall 2015, to the establishment of the NOMAD European Centre of Excellence (CoE) [19], which brought together eight research groups and four high-performance computing (HPC) centres. In order to guarantee the sustainability of the infrastructure, the FAIR-DI non-profit association [20] has been founded in association with institutions in Germany and the Netherlands with the Repository (raw data), the Archive (normalized data), and the Encyclopedia (graphical user interface for the presentation of the data) as pillars which have been developed under the umbrella of the NOMAD CoE. Recently, the NOMAD Pillar of FAIR-DI was accepted as a GO FAIR Implementation Network [21].

The basic idea of the NOMAD Repository is to enable extensive and FAIR sharing of scientific data in the computational materials science community. To that end, the NOMAD repository requests full input and output files as well as detailed information about the computer program that was used. Uploads and sharing are not restricted to a single computer code or a community, but aims at including all relevant programs and aims at involving all researchers of the field. NOMAD, and the involved computing centres, guarantee maintenance of service and data storage for at least 10 years after the last upload. Data become open access either immediately after the upload or after an embargo period of at most three years. During the embargo, data sets can be shared with selected people, e.g. collaborators or referees. Content uploaded to NOMAD is made publicly available under the Creative Commons Attribution 3.0 License (CC BY 3.0), in fact almost all current data in NOMAD are already open access. Users can curate data sets for which NOMAD issues DOIs in order to make them citable, e.g. alongside a journal publication. While uploaders need to be registered in order to also establish provenance of the data, search and download in the Repository is possible without registration.

Besides collecting, hosting, and providing data created by labs from all over the world, NOMAD is also cleaning and normalizing data. This makes the data in the NOMAD Archive independent of the employed simulation program. The required unified metadata was and still is developed in close contact with the wider community, to allow for an unambiguous labelling of the data [22]. This opens the treasure trove of computational materials science to data mining using established as well as novel artificial intelligence methods in order to identify structure, trends, correlations, and novel information out of materials big data. Ultimately, this data can be used to discover materials with new properties that may help to tackle pressing challenges of mankind, for example the more efficient use of energy.

NOMAD started as a widely FAIR compliant data service before the term FAIR had been coined. In this community, the R may also be interpreted as “re-purposable”, meaning to use data in a way not initially intended, for example looking for materials properties that were not of interest to the creator. This aspect can be also seen as one of the main drivers towards a change of spirit in this community towards FAIR and open scientific data.

5. CONCLUSIONS

These few examples from the Max Planck Society show that research organisations devoted to cutting edge research basically in all research domains took steps to optimise data management and stewardship to be able to address new scientific questions. In all cases – and there are many other examples in other research organisations – data in the order of at least Petabytes had to be integrated using harmonising methods to create a data space that is ready to be exploited to get new insights. The typical challenges of fragmentation, bad quality and also social differences had to be overcome. In all three cases, well-managed repositories that are driven by the scientific needs and harmonisation principles were the core pillars. These harmonisation principles were defined by the corresponding communities. It is not surprising that these principles are very much aligned with what have now become the FAIR principles. As Wittenburg and Strawn [23] stated, there is a clear trend towards convergence. The use of PIDs[3], agreed rich metadata, an agreement on formats and explicitness of schemas and vocabularies is key.

It should be noted, however, that the appearance of the FAIR principles has an extremely positive effect even in these cases where the basic project principles were agreed beforehand. The FAIR principles confirm the correctness of earlier decisions and their clear formulation identified the gaps which the projects need to address. In particular, the request to make all aspects machine actionable indicates where the projects need to focus their work in the future.

AUTHOR CONTRIBUTIONS

All authors have made meaningful and valuable contributions in revising and proofreading the resulting manuscript. Peter Wittenburg (peter.wittenburg@mpcdf.mpg.de) has led the editorial process of the paper submitted.

REFERENCES

[1] M. D. Wilkinson, M. Dumontier, I. Jan Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg et al…& B. Mons. The FAIR Guiding Principles for scientific data management and stewardship. Nature 3 (2016), 160018. doi: 10.1038/sdata.2016.18.
[2] EOSC. Available at: https://de.wikipedia.org/wiki/EOSC.
[3] Core Trust Seal. Available at: https://www.coretrustseal.org.
[4] IMDI. Available at: https://tla.mpi.nl/imdi-metadata/.
[5] CoC. Available at: http://dobes.mpi.nl/ethical_legal_aspects/DOBES-coc-v2.pdf.
[6] IPCC Report. Available at: https://www.ipcc.ch/assessment-report/ar6/.
[7] CMIP6. Available at: https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6.
[8] DataCite. Available at: https://datacite.org/.
[9] DublinCore. Available at: http://dublincore.org/.
[10] ISO 19135. Available at: https://en.wikipedia.org/wiki/ISO_19135.

[3] In all cases Handles and DOIs are use, both being resolved by the same globally available Handle System.

[11] Data Cite. Available at: https://schema.datacite.org/.
[12] B2FIND. Available at: https://eudat.eu/services/b2find.
[13] WDCC GUI. Available at: https://cera-www.dkrz.de/WDCC/ui/cerasearch/.
[14] Terms of Use. Available at: https://cera-www.dkrz.de/WDCC/ui/cerasearch/info?site=termsofuse.
[15] WDCC Formats. Available at: https://cera-www.dkrz.de/docs/DKRZ-LTA-Formats.pdf.
[16] CERA DB. Available at: https://www.dkrz.de/up/systems/cera.
[17] C. Draxl and M. Scheffler, NOMAD: The FAIR Concept for Big-Data-Driven Materials Science. MRS Bulletin 43 (2018), 676-682.
[18] NOMAD COE. Available at: https://nomad-coe.eu.
[19] GO FAIR IN. Available at: https://www.go-fair.org/implementation-networks/overview/.
[20] FAIR-DI. Available at: https://fairdi.eu/.
[21] NOMAD Metadata. Available at: https://www.nomad-coe.eu/the-project/nomad-archive/archive-meta-info.
[22] P. Wittenburg & G. Strawn. Common Patterns in Revolutionary Infrastructures and Data. doi: 10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0.