1st Conference on Research Data Infrastructure

Poster presentations II

https://doi.org/10.52825/CoRDI.v1i.389

Published: 07 Sept. 2023

Monitoring the state of open and FAIR data in Helmholtz

A data-harvesting and dashboard-approach by HMC

Gabriel Preuß [1][2][https://orcid.org/0000-0002-3968-2446],
Alexander M. Schmidt [1][2][https://orcid.org/0009-0005-1368-6114] ,
Mojeeb R. Sedeqi [1][2][https://orcid.org/0000-0002-9694-0122],
Vivien Serve [1][2][https://orcid.org/0000-0001-9603-7630],
Oonagh Mannix [1][2][https://orcid.org/0000-0003-0575-2853] , and
Markus Kubin [1][2][https://orcid.org/0000-0002-2209-9385]

[1]Helmholtz Metadata Collaboration (HMC), Hub Matter

[2]Helmholtz-Zentrum Berlin für Materialien und Energie, Germany

Abstract: In this contribution we present an integrated approach to monitoring and assessing the state of open and FAIR data in the Helmholtz Association. The project is part of a multi-method approach by Hub Matter in the Helmholtz Metadata Collaboration (HMC).

In a harvesting-approach, data published by Helmholtz researchers is found starting from literature metadata, harvested from the research centers. Data publications linked to that literature are identified using the SCHOLIX API. In a first approach to automated FAIR assessment, we adopted the F-UJI framework, as developed by the FAIRsFAIR consortium.

The information collected is presented in an interactive dashboard. It allows to explore in which repositories Helmholtz researchers make their data publicly available, to engage Helmholtz communities, and to identify gaps towards improving the FAIRness of Helmholtz data.

The dashboard is publicly available on https://fairdashboard.helmholtz-metadaten.de. The general approach as well as all program code are reusable by all research communities.

Keywords: OPEN DATA, FAIR DATA, METADATA HARVESTING, DASHBOARD

1 Introduction

The Helmholtz Metadata Collaboration (HMC) platform was launched in late 2019 to turn FAIR (Findable, Accessible, Interoperable, Reusable)[1] data practices within the Helmholtz Association into reality. By leveraging the visibility and re-usability of data the HMC platform aims to develop and consolidate community-expertise in metadata across Helmholtz.

To develop effective strategies, key information about the state of FAIR data practices is required:

1. Where (in which repositories) and how much is Helmholtz data published?
2. How can we identify and analyze gaps in the FAIRness of this data?

Answering these questions will help us monitor the data publishing landscape in Helmholtz and to identify action items towards a FAIR data space in Helmholtz.

2 How can we find Helmholtz data? An approach to data-harvesting and automated FAIR assessment

Being interested in finding data publications by Helmholtz researchers, we started by collecting literature metadata from Helmholtz libraries. This is done via the *OAI Protocol for Metadata Harvesting* (OAI-PMH)[2]. Data publications are often published together with them, so we can benefit from their well curated and rich metadata. The connection between literature publications and their related data publications is done via the external tool *ScholeXplorer*[3], which finds related data publications for literature publications by a DOI and offers structured metadata about them. *ScholeXplorer* allows to quickly identify and gather information about related data publications that may not be immediately apparent through other means. Finally, the toolbox evaluates data publications using F-UJI and provides a FAIR score. F-UJI [4], [5] is a web service that programmatically assesses the FAIRness of research data objects based on metrics developed by the FAIRsFAIR project[6].

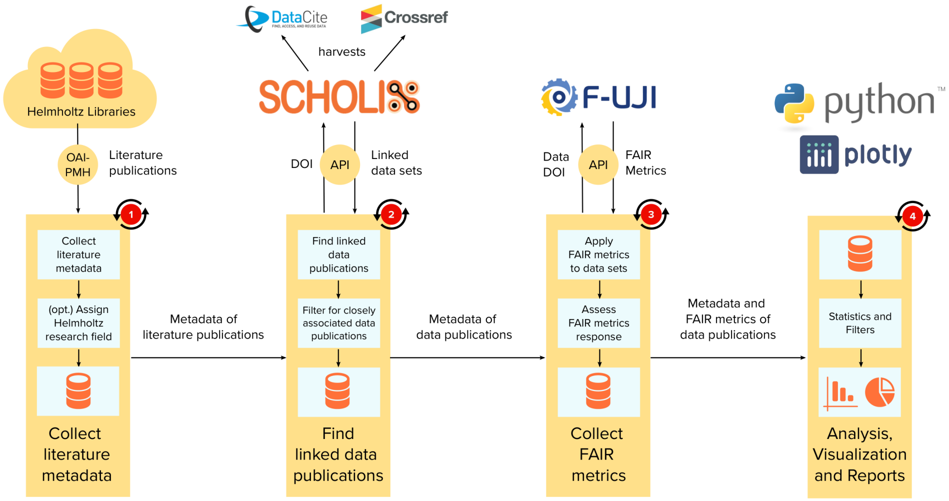

DataCite
Crossref
harvests
Helmholtz Libraries
SCHOLIX
F-UJI
python
plotly
OAI-PMH
Literature publications
DOI
API
Linked data sets
Data DOI
API
FAIR Metrics
1
2
3
4
Collect literature metadata
(opt.) Assign Helmholtz research field
Collect literature metadata
Metadata of literature publications
Find linked data publications
Filter for closely associated data publications
Find linked data publications
Metadata of data publications
Apply FAIR metrics to data sets
Assess FAIR metrics response
Collect FAIR metrics
Metadata and FAIR metrics of data publications
Statistics and Filters
Analysis, Visualization and Reports

Figure 1. Overview Toolbox Workflow

2.1 HMC Toolbox for Data-Harvesting

The implementation of the *HMC Toolbox* was done in *Python3*[7] and is based on five independently maintainable modules; named *Harvester*, *Metadata Extractor*, *Linked Data Finder*, *FAIR Meter* and the *Exporter*.

1. Harvester: the Harvester module requests a given Helmholtz library API to get information about available literature publications.
   We implemented harvesters for *OAI-PMH* APIs via the well known libraries *OAI-PMH Harvest*[8] and *Sickle: OAI-PMH for Humans*[9].
   Also, a harvester for plain file downloading is offered due to the fact that some

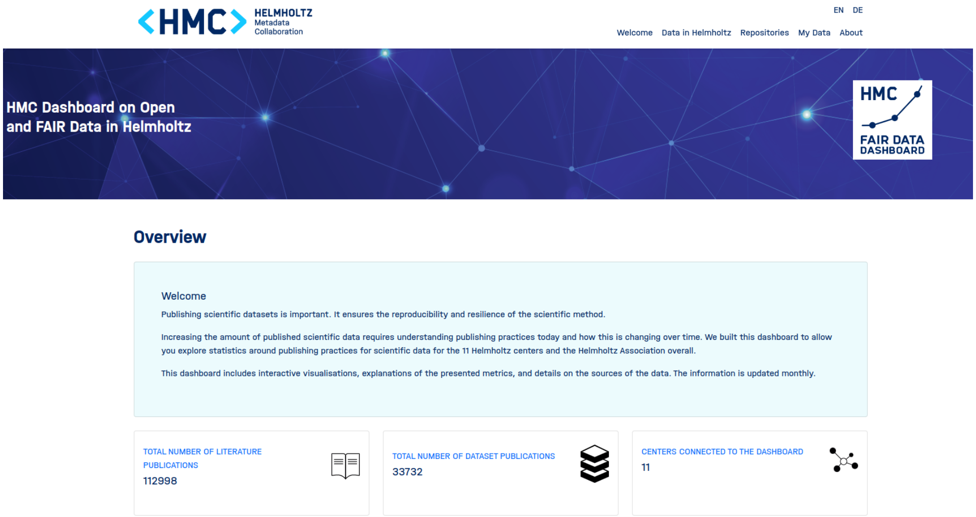
libraries offer their data in CSV files. The Harvester module offers the option to be extended by other harvester types easily.

2. Metadata Extractor: the libraries offer their data in many formats and schemas hence some kind of mapping needs to be done. Because the Dublin Core Metadata Schema is a must-have for OAI-PMH APIs the toolbox brings a mapping for that by default as well as for the old but still widely used marcxml schema.
3. Linked Data Finder: within this module the extracted metadata is enriched with additional information which can come from any implemented source. By default, the mentioned *ScholeXplorer*[3] is used to enrich a literature publication which its data publications. But also any other source is conceivable here.
4. FAIR Meter: if datasets are found for a literature publication they are successively evaluated for their compliance with the FAIR principles using F-UJI[4] as a first approach. We employ the F-UJI docker container offered on GitHub[10] to set up a local F-UJI server.
5. Exporter: to ensure re-usability of the collected data we export all metadata harvested in the *JSON* format. The Exporter module offers standard output options like screen or file output and is easily extendable to other output targets like databases.

3 How can we engage communities? An interactive dashboard approach

The information collected is presented in an interactive dashboard. It allows to explore in which repositories Helmholtz researchers make their data publicly available, to engage Helmholtz communities, and to identify gaps towards improving the FAIRness of Helmholtz data.

3.1 HMC FAIR Data Dashboard Implementation



Overview

Welcome

Publishing scientific datasets is important. It ensures the reproducibility and resilience of the scientific method.

Increasing the amount of published scientific data requires understanding publishing practices today and how this is changing over time. We built this dashboard to allow you explore statistics around publishing practices for scientific data for the 11 Helmholtz centers and the Helmholtz Association overall.

This dashboard includes interactive visualisations, explanations of the presented metrics, and details on the sources of the data. The information is updated monthly.

TOTAL NUMBER OF LITERATURE PUBLICATIONS
112998

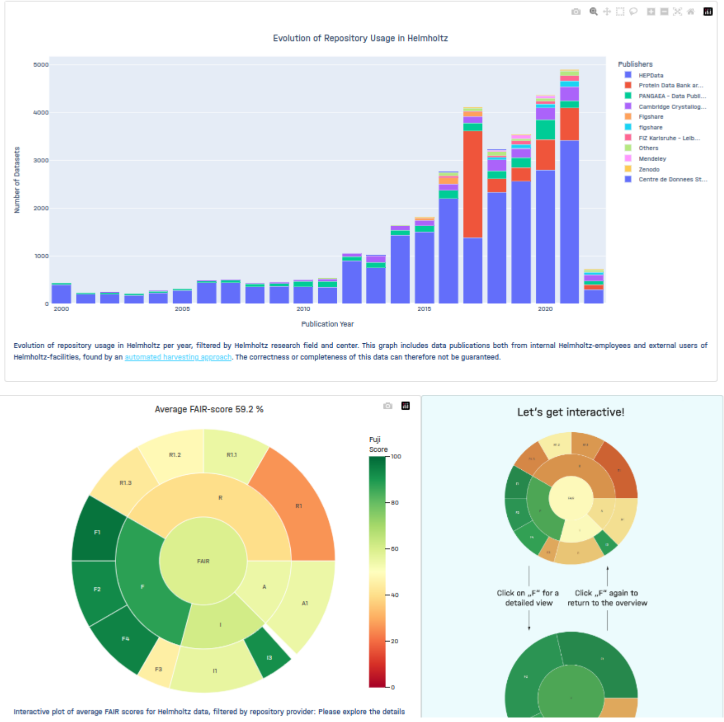
TOTAL NUMBER OF DATASET PUBLICATIONS
33732

CENTERS CONNECTED TO THE DASHBOARD
11

Figure 2. Welcome Page HMC FAIR Data Dashboard

We decided to use Dash[11] together with Flask[12] as a framework to render our plots on the dashboard.

The data collected with the *HMC Toolbox* was stored in a database and delivered to the dashboard. All plots are interactive, allowing users to filter and focus on desired issues. Additionally an option is offered to check the FAIR score of publications in the database.



Evolution of repository usage in Helmholtz per year, filtered by Helmholtz research field and center. This graph includes data publications both from internal Helmholtz-employees and external users of Helmholtz-facilities, found by an automated harvesting approach. The correctness or completeness of this data can therefore not be guaranteed.

Interactive plot of average FAIR scores for Helmholtz data, filtered by repository provider: Please explore the details

Click on „F" for a detailed view

Click „F" again to return to the overview

Figure 3. Screenshot of interactive dashboard figures

Automated approaches are limited and depend heavily on the data quality of their sources. On the other hand, the goal is to improve data quality from the FAIR perspective in the research environment. This is a mutually reinforcing loop — as the state of data is publicly communicated by the dashboard, the data producers feel empowered to improve their data which improves the quality of the harvested data for the dashboard.

The used FAIR scoring method used by F-UJI is just one perspective on the FAIRness of data publications. Other methods need to be added to the FAIR meter and evaluate the data in different ways. Additionally counting data publications is also sensitive as it can be challenging to determine what counts as a data publication. Finally there is a lack of standardized (meta-)data quality, which can make it difficult to compare results across different sources.

4 Conclusions & Outlook

We are committed to improving the data-harvesting and dashboard approaches. As a first step, the code is publicly available (see data availability statement) so that other researchers and institutions can use and build upon it. More Helmholtz centers will be integrated in later updates.
In addition, the modular approach allows for easy incorporation of other scoring metrics and harvesting options.

In conclusion, the presented approach represents an important step towards achieving a more FAIR data publishing environment in the Helmholtz research environments. By monitoring data publishing practices, and providing analyses of the data, HMC is working towards a FAIR data space in the Helmholtz Association.

Data availability statement

The HMC Dashboard on Open and FAIR data in Helmholtz is published on https://fairdashboard.helm metadaten.de/.

All program code, both for the harvesting toolbox and the interactive dashboard are made publicly available and reusable in the following GitLab repository.

Competing interests

The authors declare that they have no competing interests.

Funding

This publication was supported within the hub Matter at the Helmholtz-Zentrum Berlin by the Helmholtz Metadata Collaboration (HMC), an incubator-platform of the Helmholtz Association within the framework of the Information and Data Science strategic initiative.

Acknowledgements

We thank the Helmholtz-Zentrum Berlin for the necessary computing ressources to implement that project.

References

[1] M. Wilkinson, M. Dumontier, I. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Sci Data 3, 160018*, 2016. DOI: 10.1038/sdata.2016.18.

[2] C. Lagoze, H. V. de Sompel, M. Nelson, and S. Warner. "The open archives initiative protocol for metadata harvesting - v.2.0." (2015), [Online]. Available: https://www.openarchives.org/OAI/openarchivesprotocol.html (visited on 04/25/2023).

[3] A. Burton, H. Koers, P. Manghi, M. Stocker, *et al.*, "The scholix framework for interoperability in data-literature information exchange," *D-Lib Magazine*, vol. 23, 2017. DOI: 10.1045/january2017-burton. [Online]. Available: https://www.dlib.org/dlib/january17/burton/01burton.html.

[4] A. Devaraju and R. Huber, *F-uji - an automated fair data assessment tool*, version v1.0.0, Oct. 2020. DOI: 10.5281/zenodo.4063720. [Online]. Available: https://doi.org/10.5281/zenodo.4063720.

[5] A. Devaraju and R. Huber, "An automated solution for measuring the progress toward fair research data," *Patterns*, vol. 2(11), 2021. DOI: 10.1016/j.patter.2021.100370.

[6] A. Devaraju, R. Huber, M. Mokrane, *et al.*, *Fairsfair data object assessment metrics*, version 0.5, Apr. 2022. DOI: 10.5281/zenodo.6461229. [Online]. Available: https://doi.org/10.5281/zenodo.6461229.

[7] "Python.org, The official home of the python programming language." (), [Online]. Available: https://www.python.org/ (visited on 04/25/2023).

[8] "Github, Bloomonkey/oai-harvest: Python package for harvesting records from oai-pmh provider(s)." (), [Online]. Available: https://github.com/bloomonkey/oai-harvest (visited on 04/25/2023).

[9] "Github, Mloesch/sickle: Sickle: Oai-pmh for humans." (), [Online]. Available: https://github.com/mloesch/sickle (visited on 04/25/2023).

[10] "Github, Pangaea-data-publisher/fuji: Fairsfair research data object assessment service." (), [Online]. Available: https://github.com/pangaea-data-publisher/fuji (visited on 04/25/2023).

[11] "Github, Plotly/dash: Data apps & dashboards for python. no javascript required." (), [Online]. Available: https://github.com/plotly/dash (visited on 04/25/2023).

[12] "Github, Pallets/flask: The python micro framework for building web applications." (), [Online]. Available: https://github.com/pallets/flask (visited on 04/25/2023).