Reaching the Data Space – Standard data procedures and defining responsibilities for common data elements

Emanuel Söding[1] // Andrea Pörsch[2] // Dorthee Kottmeier[3] // Yousef Razeghi[4] // Stanislav Malinovschii[1]

0000-0002-4467-642X 0000-0003-4502-6223 0000-0002-4263-4234 0000-0002-0007-630X 0009-0002-9792-6768

[1] GEOMAR Helmholtz Centre for Ocean Research
[2] Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences
[3] Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI)
[4] Helmholtz Centre for Environmental Research – UFZ

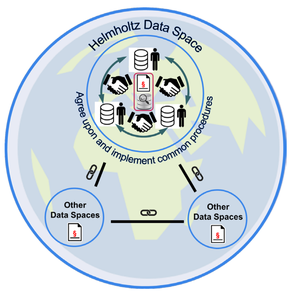
1 Introduction

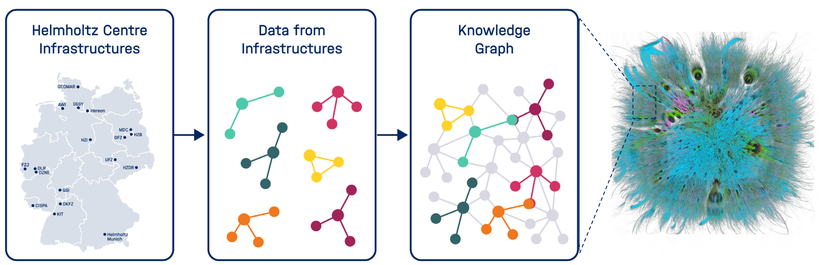
At the Helmholtz Association, we strive to establish a **well-formed harmonized data space**, connecting information across distributed data infrastructures. This requires standardizing the description of data sets with suitable metadata to achieve interoperability and machine actionability.
One way to draw logical connections between datasets and to avoid redundancy in metadata is the **consistent use of Persistent Identifiers** (PIDs). Through this method, metadata describing people, organizations, publications, vocabularies, samples, and instruments can commonly and identifiably be referenced across infrastructures. This can create a high level of interoperability allowing to build connections between data sets from different repositories according to common meta information.
In HMC we started this process by evaluating and **recommending various PIDs** for use in our data infrastructures. This, however, results in a number of changes in roles and responsibilities within our data workflows. Some of them are described here.

2 What is the FAIR data space?

The FAIR Data Space is a "**decentralized infrastructure** for trustworthy data sharing and exchange in **data ecosystems** based on **commonly agreed principles**" (Nagel L., Lycklama D., 2021).

The Helmholtz dataspace therefore consists of harmonized data infrastructures, logically tied together and aligned into a knowledge graph. This logic will allow us to connect to other graphs and data spaces and to apply advanced scientific methods to our data.





Data providers are located in the Helmholtz centres. The data records are harvested via the web. Records are consolidated and connected into a Knowledge graph. The current prototype consists of data gathered from 3 major data repositories, Helmholtz libraries and public GitLab projects (https://search.unhide.helmholtz-metadaten.de/). The consolidated graph data is quite unstructured which reflects the currently incoherent use of metadata throughout Helmholtz.

3 Some recommendations for good practice in data handling

We recommend to use the following PID systems to identify common redundant data in data infrastructures:

- **ORCID** for people and contributors to resources.
- **ROR** for organizations.
- **PIDInst** for instruments / sensors / measuring devices.
- **IGSN** for any kind of samples.
- **DataCite DOI** for data publications
- **Crossref DOI** for journal publications.

Further agreements are necessary to refer to e.g.

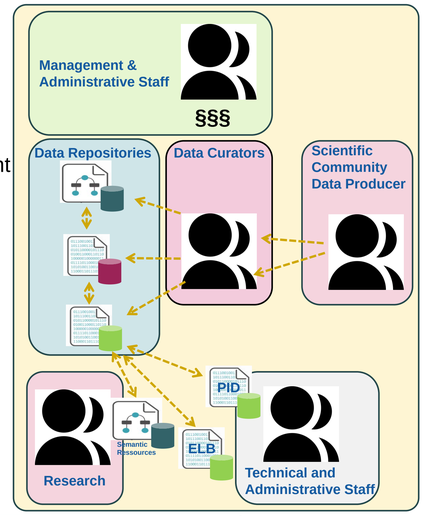
- **standardized interfaces** to harvest and exchange data
- **semantic resources** e.g. vocabularies, ontologies.
- **data containers**, e.g. FDOs or data crates, to achieve machine actionability

These recommendations require special responsibilities for particular stakeholder roles in the organization.

4 Roles

Data maintaining roles:

- Organization Management
  - contributors
  - technicians
  - administration
- scientists producing data
- Data curators
- Data infrastructures / repositories



5 What is YOUR Responsibility?

Organization:

- Employ measures and incentives to **encourage all employees** to register with ORCID and to keep their metadata current.
- **Keep record** of your staffs ORCIDs and share them with the RDM teams.
- **make a person or unit responsible** to maintain the centre's ROR record.
- keep the ROR entry **up-to-date**.
- **publish your ROR** in an appropriate place(s), e.g. the imprint of your website.
- employ measures to **register instruments with PIDInst at acquisition**.
- make a person responsible to **maintain the PIDInst** for each instrument (lab techn's).
- make a person or unit responsible to **maintain the centre's IGSNs**.

Contributor to resources:

- **register your ORCID** if not yet done
- keep your **ORCID metadata current**
- **share** the relevant data with your center.
- **publish your ORCID** as part of your contact details, e.g. in your email signature

Technician:

- **register a PIDInst** for instruments not yet registered.
- keep the **PIDInst metadata current**.
- put a **written record** of the PIDInst on every instrument, where possible
- record the PIDInst with **measured data set**.

Scientist producing data:

- **record ORCIDs** with persons mentioned – *look up in ORCID directory*
- **record RORs** with organizations mentioned – *look up on imprint of website*
- **record PIDInsts** with any measurement *- you should find it on the instrument.*
- **record an IGSN** with any sample taken.
- record the **parent IGSN** with subsamples

Data curator:

- **enable, train and encourage** staff to register IGSNs when samples **are taken**.
- **enable, train and encourage** staff to record parent IGSNs with **subsamples**.

Data infrastructure:

- record an ORCID with any person
- treat **ORCID** metadata as the **primary source of truth** and update your own metadata accordingly.
- **inform persons** about inaccurate ORCID metadata, or request permission to update.
- **record a ROR** to identify organizations and make this data **available for harvesting**.
- treat **ROR** metadata as the **primary source of truth** and update your own metadata accordingly.
- **inform organizations** if the metadata registered with the ROR is not accurate.
- record a **PIDInst with instruments** registered in data infrastructures.
- treat **PIDInst** metadata as the **primary source of truth** and update your own metadata accordingly.
- **inform the responsible person** if you think the PIDInst metadata is not accurate.
- **record an IGSN** to identify samples and parent samples and make this data part of the metadata **available for harvesting**.
- treat **IGSN** metadata as the **primary source of truth** and update your own metadata accordingly.

Further Information and Contact:

HMC Earth and Environment
E-mail: hmc-hub-ee@geomar.de
www.helmholtz-metadaten.de