HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

HMC PAPER | 1 | FAIR PRINCIPLES

An interpretation of the FAIR principles to guide implementations in the HMC digital ecosystem

February 2022

Short abstract

Findable, accessible, interoperable and reusable (FAIR) set principles that determine best practice for managing the dissemination and ensuring longevity of digital resources. The Helmholtz Metadata Collaboration (HMC) provides guidance on metadata and related topics to those working in the Helmholtz ecosystem. Given the complexity - both of the FAIR principles, and the Helmholtz ecosystem - we interpret the principles so they can be directly applicable to the Helmholtz context. In this interpretation we consider managers, tool-developers, data managers, and researchers amongst others; and provide guidance to these disparate roles on applying the FAIR principles in their professional lives.

Keywords: FAIR, HMC

DOI: tbd

Citation: Helmholtz Metadata Collaboration; An interpretation of the FAIR principles to guide implementations in the HMC digital ecosystem, 2022.

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

Call for Review

You are all invited to comment on this version. Please send your feedback by email to hmc-info@geomar.de.

Version: 1.0

This document was generated in a FAIR manner. All previous versions are available on request.

Authors (ORCID)**:** Pier Luigi Buttigieg (0000-0002-4366-3088), Constanze Curdt (0000-0002-9606-9883), Ahmad Z. Ihsan (0000-0002-1008-4530), Thomas Jejkal (0000-0003-2804-688X), Markus Kubin (0000-0002-2209-9385), Oonagh Mannix (0000-0003-0575-2853), Daniel P. Mohr (0000-0002-9382-6586), Anton Pirogov (0000-0002-5077-7497), Barbara Port (0000-0002-8941-954X), Karl-Uwe Stucky (0000-0002-0065-0762)

HMC group: Cross-cutting Topic Working Group “FAIR Concepts and Implementation”

Contact: HMC Office
GEOMAR Helmholtz Centre for Ocean Reaearch Kiel
Wischhoftstr. 1-3
24148 Kiel, GERMANY
E-mail: hmc-info@geomar.de

www.helmholtz-metadaten.de

Content

1 Preamble 4
2 Actions moving forward 5
3 Definitions 7
4 Findability 9
5 Accessibility 12
6 Interoperability 15
7 Reusability 18
8 Assumptions, warnings, and caveats 22
9 Application case snippets (incomplete/unreviewed) 22
Findable 22
Accessible 23
Interoperable 23
Reusable 23
References 25

1 Preamble

Within the Helmholtz Association, the Helmholtz Metadata Collaboration (HMC [1]) is part of the Helmholtz Incubator Framework for Information and Data Science. Its mission is to facilitate the discovery, access, machine-readability, and reuse of research data across and beyond the Helmholtz Association. This document, compiled by an HMC cross-cutting topic team (CCT3, see Definitions), provides initial guidance on the interpretation of the FAIR guiding principles in the Helmholtz context.

The FAIR guiding principles [2] (published by Wilkinson et al. 2016 [3]) provide high-level advice on making digital resources Findable, Accessible, Interoperable, and Reusable. These principles summarise a body of long-standing best practices for sharing data via both internal networks and the internet.

Mid- and low-level interpretations of the FAIR principles are needed to adapt them to specific usage scenarios. Generalised interpretations have been provided by groups such as GO FAIR [4] and the RDA (e.g. the RDA's FAIR Maturity Model [5]), as well as institutes such as the SNF [6] (Swiss National Science Foundation); guidance which this document drew from in order to develop these recommendations to the Helmholtz community. The Helmholtz/HMC usage scenario is a complex one, wherein the multidisciplinary perspectives of individual groups, sections, institutes, Hubs, and digital infrastructures of the Helmholtz Association (e.g. HMC, the Helmholtz Artificial Intelligence Cooperation Unit [7], and the Helmholtz Imaging Platform (HIP [8])) must be synthesised across research and operations. This synthesis must also be outward looking, in that it natively interoperates with regional and global systems.

The scale of this endeavour is formidable. This document – composed by representatives of the HMC's thematic Hubs, work packages, and other organisational units organised in a cross-cutting topic working group – aims to nucleate a more inclusive process to provide concrete guidance around the FAIR principles, supporting tool developers, data managers, (digital) technicians, researchers, and all other Helmholtz staff seeking to secure their digital legacies. Please note that as the HMC, definitions and technologies evolve this document will also evolve. We will endeavour to communicate the latest version, but be aware that new versions may appear.

2 Actions moving forward

The interpretations in this document are the basis upon which CCT3's future activities will support the HMC. Some of these activities are listed below, and will inform updated and extended versions of this document:

1. Evaluate sets of minimal metadata (e.g. used for FAIR Digital Object construction) proposed by HMC working groups against the criteria outlined in each of the FAIR principles
    - Some of these minimal sets will be used for discipline-agnostic indexing and retrieval HMC core processes (e.g. same logic as metadata on DOIs, not discipline specific), while others will likely be Hub-specific.
    - Assess the intention of each field and its fitness-for-purpose.
    - Provide clear guidance on how to avoid under- and overloading each, describing how to create compatible extensions
    - Provide feedback on how to qualify (as described below) each field using semantically described field names and relations (e.g. `rdfs:seeAlso`, `dc:license`, etc), in coordination with HMC's CCT7 (Glossary and Semantics).
2. Support activities to create Hub-specific extensions to HMC-level minimal metadata specifications, collaborating with CCT1
3. Develop recommendations concerning the versioning of data, metadata, and other digital assets to enable a structured and transparent approach for publishing revisions and corrections of research data.
4. Define an approach for provenance tracking metadata via global standards (e.g. via PROV [9]).
5. Once FAIR data exchanges begin, evaluate if the recommendations and interpretations in this document and documents generated by 3. and 4. are effectively met.
6. Support HMC working groups in developing or applying checklists for self-assessment of compliance with this document for different user groups.
7. As the need arises, partner with other CCTs to draft guidelines clarifying good (eventually best) practices for FAIR research and operational data publication, both globally and in each Hub. Collaboration themes may include:
    - CCT1 (Metadata Landscape Mapping) conducting surveys to identify sharing needs, strengths, gaps, etc. to inform our activity
    - CCT2 (Training and Outreach) helps our user community to select and apply good practices by providing training materials and workshops

- CCT4 (From Development to Deployment - Software, Tools, Workflows and Interfaces) supplying tooling to support the creation, validation, and publishing of digital assets
- CCT5 (Central Community Services) collecting publishing guidance and related information from CCT3 and providing a platform to communicate this with the wider community
- CCT6 (Communication) creation and execution of a targeted dissemination strategy to relevant stakeholders
- CCT7 (Glossary and Semantics) to ensure terminology usage is consistent and explicitly defined (and eventually machine-actionable) during our activities

3 Definitions

Note: The HMC is developing a semantically consistent glossary whose content and definitions will supersede those below in future versions of this document. The definitions below are for informal orientation and guidance.

Application: a) in software, a program that performs a set of tasks b) more generally, the act of using something for a defined purpose.

Benefit: A positive impact upon a given entity.

Communication protocol: "A communication protocol is a system of rules that allow two or more entities of a communications system to transmit information via any kind of variation of a physical quantity. The protocol defines the rules, syntax, semantics and synchronisation of communication and possible error recovery methods. Protocols may be implemented by hardware, software, or a combination of both." [10]

Cross-cutting topic (CCT): A topic identified as relevant across the thematic Hubs and work packages (or other organisational units) of the Helmholtz Metadata Collaboration (HMC). Each CCT is championed by a team from across the HMC to ensure wide representation of perspectives and expertise. CCT teams / working groups are convened and adjourned dynamically, as needs arise.

Data: A quantitative or qualitative representational entity encoded as a sign, symbol, marking, value, or pattern on any medium.

General user: In this context a general user is any human or machine agent interacting with HMC or Helmholtz digital services. In the case of a human agent, sufficient computer literacy and domain knowledge to access and use these services is assumed; but no advanced programmatic or computational skills. In the case of machine agents, capacity to use generic web interoperability standards (e.g. W3C [11]), and similar standards within a community of practice/expertise are assumed. Please note that "sufficiency" as used in this definition is variable and depends on the complexity of a particular service.

Information: A datum or data which, when accurately interpreted, reduces uncertainty about the properties or behaviours of an entity.

Knowledge: A representational entity which 1) is an abstraction of an entity constructed from information about that entity, 2) grants its bearer reliable familiarity with that entity, and 3) can be used to reason about that entity.

Metadata: Data which is about other (meta)data.

Metadata, intrinsic: Metadata which describes a property of the (meta)data they describe (e.g. its size, format, or other metadata often generated by the processes creating (meta)data itself).

Metadata, contextual: Metadata which describes external processes, factors, or other entities which provide information about how or in what kind of environment the (meta)data they describe were generated, modified, or otherwise processed.

Metadata, scientific-grade: Metadata which provide sufficient detail (e.g. high-quality provenance data, links to methodology, links to calibration data) for independent parties to 1) reproduce the planned processes (i.e. the intentional application of the scientific method) which generated the data they describe as well as 2) understand unplanned events which influenced the results. Roughly, a third party should know what was intentionally done to generate the (meta)data, and if anything, unexpected or uncontrollable happened to ensure reproducibility.

Resource: Any tangible or intangible entity which is capable of conferring benefit to its user(s).

Rich metadata: Metadata which is complete, accurate, and standardised enough to allow a wide range of operations defined by a user community or within an application case. Note: concretising this definition has been identified by CCT3 as a task to accomplish with further input across HMC applications. Further note that richness in one FAIR principle does not mean richness across all (can be rich in F but poor in I).

Stream, Data: Continuously transferred data (i.e. streaming data). Note that data streams are often collected – in part or in whole – as a collection or data set.

User: A human or machine which accesses, interacts with, and attempts to obtain benefit from a resource.

4 Findability

The first step in (re)using data is to find them. Data should be easily findable for both humans and computers, through both specialised and generic search and discovery interfaces. Further, the metadata associated with findable data will support the findability of related data or data sets using qualified references (see Interoperability). Thus, machine-readable and standardised metadata are essential for efficient accurate retrieval and automatic discovery of data sets and services.

The Findability principle has the following criteria:

F1 (Meta)data are assigned a globally unique and eternally persistent identifier.

F2 Data are described with rich metadata.

F3 Metadata clearly and explicitly include the identifier of the data it describes

F4 (Meta)data are registered or indexed in a searchable resource.

Regarding **F1**: (Meta)data are assigned a globally unique and eternally persistent identifier.

- The HMC will provide recommendations and support for (meta)data (be it in static objects, dynamic objects, or streams) and other digital resources (including services and software) to be assigned persistent and globally unique identifiers.
- Eternal persistence cannot be assured, but the spirit of the principle is noted, and CCT3 strongly recommends that only services that have a multi-decadal sustainability plan should be used to issue authoritative identifiers.
- The HMC will likely use a mix of identifiers that are issued and maintained by both HMC facilities and external services.
- Once identifiers are published, they must become dereferenceable over the web within a short time (maximally, days) and remain resolvable over the web permanently (excluding unavoidable and inevitable service disruptions, like those during data migrations).
- On demand, the HMC will provide context-sensitive guidance to partners in identifying both where and how a globally unique and persistent identifier can be acquired. It is likely that it will be via web-UI to catalogues and services which can issue identifiers and to retrieve existing ones.
- The HMC will support users – through services or tools – in discovering and tracking which identifiers are associated with Helmholtz (meta)data.

Regarding **F2**: Data are described with rich metadata (cf. **R1**).

- Rich metadata (see definition, above) include application-specific descriptors, for example: names, locations (geolocations), funder metadata, campaign or study design data, (persistent) identifiers of objects and samples investigated.
- Rich metadata allow a user to find data based on multiple query types, targeting their provenance, properties (e.g. size, formats, quality control) and upstream generation, processing, and storage of the (meta)data they describe.
- For entities which change their state after (meta)data has been produced (e.g. a physical sample undergoing destructive sampling), rich metadata will allow a user to understand (and if possible replicate) the state of an entity before, during, or after (meta)data collection.
- In the Helmholtz digital ecosystem, rich metadata will support findability through queries directed and discovery paths related to scientific (re)publication, data management, and strategic planning.
- In order to determine the minimal threshold for richness, the HMC will evaluate needs across its Hubs, capturing the typical search and discovery paths of data through their metadata in support of research and operational activities.
- The “richness” of metadata is typically tied to discipline-specific approaches. The HMC will foster the discussion and consolidation of these approaches across the Helmholtz digital ecosystem, and synthesise guidance accordingly. Initially, the HMC’s CCT mechanism (see Definitions) will be used to pursue this goal:
  - CCT1 will engage research data repositories relevant to Helmholtz stakeholders and work with CCT3 to evaluate their understandings of rich metadata.
  - CCT1 will catalogue tools and approaches which support users in augmenting research data with metadata, including – with CCT3 – evaluation of richness of metadata for major application types.
  - CCT2 will support the users in selecting and applying these tools by providing training (materials) and workshops with respect to discipline-specific best practices in how to augment research data with rich metadata.
  - CCT5 will support users in discovering and selecting this training material together with the tools and approaches. This will enable to user to discover and select relevant information and enable the HMC to harvest the needs of the community.

Regarding **F3**: Metadata clearly and explicitly include the identifier of the data it describes

- Findability relies on the unique identifiers described in **F1** being discoverable and intelligible to both machines and humans.
- Identifiers must be in fields explicitly described by globally standardised (and therefore generically discoverable) field names, appropriate to the format and schema being applied (e.g. an `owl:NamedIndividual`, or a `schema.org` identifier associated with a Thing, or the primary Identifier attribute of a DataCite record).

Regarding **F4**: (Meta)data are registered or indexed in a searchable resource.

- Helmholtz personnel use a diverse collection of both internal and external searchable resources to register and cross-index their (meta)data.
- In some cases, data will not be registered or indexed (e.g. due to confidentiality, sensitivity, embargo), but their contextual or intrinsic metadata may be, to allow findability and contact with the responsible parties listed in their provenance fields.
- HMC's FAIR Data Commons (through Task 2.1.2 of the HMC Work Plan) will generate tools to support Hubs in generating indices and registries, searchable for metadata and capable of interoperating with one another and any central HMC technologies, including the FAIR Digital Object architecture (Task 2.1.1).

5 Accessibility

Once found, (meta)data must be accessible for further use. That is, once a dereferenceable URI, IRI, or other identifier has been retrieved via a search or discovery routine, there must be some supporting architecture and protocol to access the (meta)data it points to. In modern contexts, access is almost always secured through the internet, except when dealing with confidential data that may be stored on local systems and networks. The latter case does not nullify the need for FAIRness, but merely means that accessibility and findability may be limited in secure systems and networks.

The Accessibility principle has the following criteria:

A1 (Meta)data are retrievable by their identifier using a standardised communications protocol.

A1.1 The protocol is open, free, and universally implementable.

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary.

A2 Metadata are accessible, even when the data are no longer available.

Regarding **A1**: (Meta)data are retrievable by their identifier using a standardised communications protocol.

- When a dereferenceable identifier (cf. **F1**) to a (meta)data object or stream has been secured, one can use a standardised communication protocol (e.g. TCP/UDP-based protocols) to follow that identifier to the (meta)data it is associated with.
- Where access to data is restricted, the identifier should resolve to metadata including provenance and current contact information (an ORCID, email, telephone number, etc) of an individual or department who/which can be approached to provide access to the data.
- The HMC will identify which communication protocols are in use for Helmholtz digital systems. The HMC will offer open reference implementations of key protocols to allow other parties within Helmholtz to establish dedicated access routes to the (meta)data they manage, if needed.
- The HMC will develop recommendations and policies regarding the minimal metadata which should be accessed when following a URI or similar (see Findability).

Regarding **A1.1**: The protocol is open, free, and universally implementable.

- Anyone having an internet connection and a computer should be able to access at least the HMC-specified minimal metadata associated with a URI or similar identifier.
- This should be possible using one or more standardised high-level protocols, e.g. OAI-PMH [12], ResourceSync [13] or the Digital Object Interface Protocol [14].

- All protocols in use should be well documented, re-implementation should be possible without running into licensing issues and in any programming language. Systems should not (solely) rely on a proprietary or commercial communication protocol.
- The HMC will foster the use of open, free, and universally implementable communication protocols in Helmholtz's digital infrastructure.

Regarding **A1.2**: The protocol allows for an authentication and authorisation procedure, where necessary.

- When necessary (e.g. to protect intellectual property rights, embargoes, confidentiality, etc), the protocols used by the HMC's FAIR digital ecosystem must support user management functions to authenticate users and confirm their right to access a (meta)data record or stream.
- The ability to allocate rights on a per user basis should be supported, as well as user roles which allow the assignment of rights without central tech support (e.g. a PI granting a new PhD candidate access to a project's data store accessible via a set of URIs).
- For HMC services, the integration of a centralised Authentication and Authorisation Infrastructure (AAI), e.g. offered by HIFIS via the Helmholtz AAI, should be supported.
- Machine agents should also be subject to authentication and authorisation in order to facilitate controlled machine-actionability. These machine agents would also have user accounts, managed by Helmholtz personnel with equal or broader access rights.
- As needed, the HMC will convene task teams to draft recommendations for managing authentication and authorisation processes in a FAIR digital ecosystem.

Regarding **A2**: Metadata are accessible, even when the data are no longer available.

- Depending on funding, local policies, project turnover, or technical events, data in the Helmholtz digital ecosystem might be (re-)moved over time, e.g. to a tape archive from where it is no longer accessible without additional effort.
- It is key that the metadata associated with these digital assets is not made inaccessible by such events.
- Users and services should be able to access the metadata, or at least a commonly agreed subset of metadata, containing, e.g. contact information for the individual responsible for providing data access or tombstone information if the data has physically gone.
- A HMC's persistent identifier resolving system will provide a guarantee of long-term persistence for at least the kernel metadata associated with records with an associated FAIR Digital Object. However, this metadata intentionally minimal.

- Additional metadata required to fulfil this aspect of accessibility to the satisfaction of a disciplinary community (e.g. in an institute or across a Hub) should be provided and persistently stored by the institution which is responsible for the data. The HMC will provide, as needed, support to create a distributed, but interoperable, collection of such institutionally based metadata stores.
- Where this is not possible, policies and interfaces should allow the extraction and migration of metadata in order to preserve accessibility.

6 Interoperability

Once (meta)data have been found and accessed, a major part of their utility and value is determined by how interoperable they are (i.e., how well they "work together") with other (meta)data and software. If 1) community standards (e.g. DwC [15], NetCDF [16], or OGC standards [17]) are followed carefully and consistently and 2) those standards themselves are interoperable, the interoperability of (meta)data is increased and fewer resources are needed for (meta)data preparation and transformations prior to research or application.

The FAIR principles (and thus this document) focus on semantic interoperability (i.e. using the same vocabularies and ontologies or those which testably interoperate with one another using conventional mappings), rather than interoperability at the level of (meta)data structures, formats, types, encodings, and so forth. We recommend that additional guidance on such low-level interoperability is also produced and aligned, within and across Hubs. While they are operational, this activity will be supported by CCT3, in collaboration with CCTs 1 (Mapping), 4 (Tooling), and 7 (Semantics), in pursuit of Tasks 1.3.2, 2.2.3, 3.2.1 of the HMC Work Plan.

The Interoperability principle has the following criteria:

I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2 (Meta)data use vocabularies that follow FAIR principles.

I3 (Meta)data include qualified references to other (meta)data.

Regarding **I1**: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- Clarifications
  - Knowledge representation (KR) is a branch of Artificial Intelligence (AI) that uses methods to make human knowledge machine-readable and -actionable. KR often encodes knowledge in machine-readable terminologies (e.g. controlled vocabularies, glossaries, thesauri, ontologies) which can be combined to form "languages". To be useful in an operational sense, these are digitised and shared in formats like RDF, RDFS, RDF*, SKOS, and OWL.
  - Formal languages use constrained and machine-readable relations between terms (culminating in mathematical and descriptive logics) to allow machine agents to understand relationships between terms.
  - Shared languages are those that are used by many independent systems. Much like natural language, a lingua franca promotes shared understanding and the ability to work together.
  - Accessible languages are those that comply/align with the FAIR Accessibility principle (see Accessibility)

  - Broadly applicable languages are those that are fit for use by independent communities over a range of applications and usage scenarios. These may be composed of one or more interoperating terminologies.

- To enrich (meta)data with languages for KR, the HMC will support the Helmholtz community in coherently selecting and using terminologies to a) provide controlled and machine-actionable values for their (meta)data fields (i.e. in preference to free text) and b) be used to describe the fields themselves.
- Importantly, the HMC will encourage and support more consistent and broader use of Hub-specific and general KR resources, which its teams have vetted for quality and sustainability. This is essential to ensure that the "shared" aspect of this principle is met.
- An HMC CCT (CCT7) has been established to begin normalising semantics across HMC activities, with an intention to mature internal glossaries into ontologies and/or other KR technologies.

Regarding **I2**: (Meta)data use vocabularies that follow FAIR principles.

- The vocabularies (or other terminologies) called for in **I1** must, themselves, be FAIR. A custom vocabulary used by a group, institute, project, or even infrastructure is of little use beyond those confines unless it participates in a wider community of FAIR KR.
- The terminology must be Findable in some form of index (e.g. OntoBee [18] or the EBI OLS [19] for ontologies, and resources such as INFORMEA [20] Glossary for less-expressive resources), Accessible via TCP-IP/UDP, HTTP, RESTful and/or SPARQL endpoints, Interoperable with other terminologies, and under appropriate licensing to be reusable (ideally CC0).
- It is essential that the terminology interoperates with others both within and beyond its immediate sphere of operation (e.g. by importing, cross-linking, or mapping terms/classes in quality-controlled and machine-readable ways). If this is not met, then the resource is not necessarily useful for broad interoperation.
- Recent commentary [21] on what makes a KR resource FAIR has been released by the FAIRsFAIR consortium
- The HMC will provide additional guidance to the Helmholtz community on how to vet and select terminologies for use, as well as providing Hub-specific recommendations and commentary on existing resources. For terminologies which have wide community uptake but do not meet these criteria, the HMC will - on demand from its users - review the resource and explore the possibility of elevating its technical basis. Wide community uptake indicates a resource which is well adopted (i.e. used as a de facto or de jure standard, or a regular part of data management) in a multi-institute community of practice, nation or multi-national setting; furthermore, the resource should have a

sustainability plan beyond a few projects cycles and a team of independent maintainers from multiple institutes.

Regarding **I3**: (Meta)data include qualified references to other (meta)data.

- In a key-value example, A qualified reference minimally constitutes 1) a key defined by a well-adopted semantic resource and 2) a value which dereferences to other metadata. For example, using "[key] [value]" syntax to identify a contributor using Dublin Core:

```
dcterms:contributor <https://orcid.org/0000-0002-4366-3088>
```

- (Meta)data never occur in isolation, and findable and accessible links to other (meta)data - stored in separate records - are typically needed for the Helmholtz community (and wider world) to understand their context and provenance. In this way, metadata records link together and enrich related digital objects and streams.
- Qualification of the references that link one digital entity to another follows similar logic to **F3**, in that the qualification explains the intention of the reference. That is, the qualification explains how the reference relates to the digital entity whose metadata one is viewing. A URI to another data set can be qualified with standardised properties which have global (as opposed to community) adoption such as `rdfs:seeAlso`, `rdfs:member`, `owl:priorVersion`, `owl:incompatibleWith`, `schema:Person or`, `dc:accessRights`. Of course, unstandardized properties can easily be used, but this is strongly discouraged, as these are not broadly interpretable.

7 Reusability

Finally, once a digital resource has been found, accessed, and interoperably integrated with others, users typically wish to reuse its content.

Reusability of a resource depends on the quality and integrity of its content, as well as how well it is described, contextualised, and standardised (both internally and with reference to community standards). Aside from its intrinsic properties, a resource can only be ethically reused if there is clear provenance (clarifying how and why the (meta)data was collected) and licensing information provided in the metadata. These latter aspects are becoming increasingly well-articulated as principles and policies such as the CARE [22] emerge.

The Reusability principle has the following criteria:

R1 (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1 (Meta)data are released with a clear and accessible data usage license

R1.2 (Meta)data are associated with detailed provenance

R1.3 (Meta)data meet domain-relevant community standards

Regarding **R1**: (Meta)data are richly described with a plurality of accurate and relevant attributes

- (Meta)data without context is rarely useful, let alone reusable. Thus, only (meta)data that is accompanied with sufficient contextual metadata for a user to interpret its meaning, relevance, implications, and other aspects that elevates the data to information and contributes to knowledge.
- The level of richness needed depends on the intended application by any user. As the range of such applications is unknown, (meta)data generators should endeavour to collect as much contextualising (meta)data as feasible, following their community's best estimation of complete contextualisation (e.g. following the logic of Minimal Information Standards [23]). The metadata collected may not immediately be associated with any standard specification, however, this should not prevent its collection and - if the creators believe it to be vital for their community's work - a request to the relevant standards body to update the specification should be made.
- For scientific-grade data (see Definitions), the minimal set of descriptive (meta)data should allow reproducibility of the processes that generated that (meta)data. Thus, the metadata could include values pertaining to environmental and laboratory conditions and processing parameters, as well as links to digitised manuals, protocols, code repositories, and similar methodological elements which are sustainably archived.

- Richness does not necessarily translate to having complex (i.e. difficult for machines to understand unambiguously) (meta)data types. For example, if "Temperature" is considered a relevant attribute, the (meta)data values describing it should be as atomic as possible, disaggregating numeric values from unit strings. A key "`temperature value`" with a value of "12" associated with an XSD encoding "`decimal`", accompanied by another key "`temperature unit` with a value "`Celsius`" is less computationally complex than a compound, string/free-text value `12 C` or its variants (`12C`, `12 degrees C`, `12 * Celsius`, etc). Ideally, the latter field would be made less arbitrary by using a widely-adopted and high-quality controlled vocabulary or similar semantic resource, e.g. using Quantities, Units, Dimensions and Data Types terminology (QUDT):

```
temperature unit <http://qudt.org/vocab/quantitykind/CelsiusTemperature>
```

  Such considerations and measures will also affect interoperability and findability.

- The HMC, on request, will support Helmholtz digital stakeholders in identifying adequate levels of richness and developing solutions to store requisite (meta)data to ensure Reusability.

Regarding **R1.1**: (Meta)data are released with a clear and accessible data usage license

- The use of standard and legally sound licenses is crucial for reusability in multiple scenarios, but especially when a collections (meta)data from different sources are used. As data-intensive and synthetic science deals with ever-larger volumes and variety of data, it is generally unfeasible to ensure compatibility of those many different licenses unless they are clearly recorded in qualified metadata fields (e.g. `http://purl.org/dc/terms/license`).
- We thus **strongly recommend** that all (meta)data circulated through the Helmholtz digital ecosystem is associated with usage license, such as those listed in this licence selector [24].
  - We strongly recommend following established conventions for expressing license information, such as through Dublin Core's dcterms:license [25]. An RDF representation of a Dublin Core license expression is below, note the use of the stable URL to the license:

```
ex:mySoftware     dcterms:title "GeoNetwork - Geographic Metadata Catalog" ;
                  dcterms:license <http://www.gnu.org/licenses/gpl.html>.

<http://www.gnu.org/licenses/gpl.html> rdfs:label "GNU General Public License".
```

- In a key value pair representation, this may be cast as

```
https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#terms-
license http://www.gnu.org/licenses/gpl.html
```

- The HMC will support Helmholtz stakeholders in (further) developing and aligning policies on metadata release/publication in line with this Principle.
- The HMC will socialise the notion that ad hoc, disciplinary conventions around embargoes on open data are not equivalent to formal data usage licenses: these can and have been inadvertently violated due to their ambiguity and misinterpretation.

Regarding **R1.2**: (Meta)data are associated with detailed provenance

- The provenance – the authenticated origin and derivation history – of a digital entity is essential to securing trust in the entity's content and, thus, its role in reproducible science and operations. On a content level, the HMC will develop:
  - Guidelines on securing a minimal standard of completeness for each Hub and discipline, ensuring, for example, reproducibility of the experiments/ observations that generated the (meta)data from scientific activities
  - Guidelines on how to include the results of calibration, error calculation, and quality assessment at each step of the provenance chain
  - Guidelines on how to include diagnostic results and code/methods for each step of the provenance chain.
- The HMC will explore standards (e.g. PROV), and associated archives and tooling, to link consistent provenance information with each digital entity in its purview.
- Checksums or hashes (ideally SHA-512) of data records should be included in the corresponding metadata records along the entire provenance chain. At a higher level, checksums or hashes of (meta)data records can be placed in any PID record referencing those. Hashes of PID records and their content are generated by the PID management service used by the HMC (Handle PID system).
- Manually entering full provenance information can be prohibitive, thus, the HMC will support Hubs and institutes in streamlining, automating, and standardising their methods for creating provenance information.
- We strongly recommend that - at a minimum - rich metadata about the hardware, operating system(s), and software (including all parameter settings) used to process the data is included in the metadata relevant to **R1.2**.
- Additionally, we also strongly recommend the software itself is also archived to preserve provenance chains and ensure reusability/reproducibility. This is especially the case for academic software. We note that securing the operating system needed to run the

software (e.g. in a container) is an additional complexity as are variations introduced at the hardware level. At this time, we have no generic recommendations for these complexities, but invite Helmholtz personnel to consult with the HMC on specific cases.

Regarding **R1.3**: (Meta)data meet domain-relevant community standards

- Across most disciplines, some community-level data and/or information standards (de facto or de jure) have been established and disseminated. Adoption of quality standards within domains greatly enhances reusability and – if semantically augmented – interoperability of digital entities.
- We note that community conventions have varying quality, and may require vetting prior to integration into Helmholtz-wide systems.
- The HMC will - on request - assist Helmholtz personnel find, evaluate, or adopt well-established and high-quality standards and conventions in their field.
- In considering such standards, preference will be given to those which
  - depend only on non-proprietary, open formats, tools, and process for development and use
  - have an open and transparent process for revision and extension.
  - follow the FAIR principles to represent and release their standards, thus allowing Helmholtz users to effectively use them to qualify their associated (meta)data.
  - Are implemented using the simplest, yet fit-for-purpose, form. That is, (meta)data encodings from the record to the individual value level should not require dedicated scripts or parsing software to be understood by software capable of handling generic (meta)data exchange via, e.g., RDF, JSON(-LD), XML.
- Standards/Conventions for 1) using strings to identify variables, dimensions, or other entities and 2) using controlled terms as values of (meta)data fields are a potential overlap with the terminologies described in the Interoperability section, above.
  - If the IRIs for any terminologies used do not dereference and resolve to a resource where additional metadata can be discovered, then reference to the standards and the conventions for vocabularies/ontologies must be referenced in the metadata.

8 Assumptions, warnings, and caveats

1. **Assumption**: We assume that URIs (and similar) will resolve to (meta)data records maintained, under version control, by trusted repositories with multi-decadal sustainability. When these URIs are aggregated (e.g. in a FAIR Digital Object), they must reliably bind together the intended versions of each digital entity they reference. If the target of the URI changes (e.g. because a repository shuts down and transfers its holdings to another), the URI should stay the same, but resolve to the new location.
2. **Warning**: In the long-term (decadal scale), many repositories offering URIs to (meta)data, software, code, and other digital objects shut down without redirecting those URIs to another trusted repository which assumes their custodianship. This raises the question of whether the HMC or a related Platform offers a "deep-time" archiving service for at-risk digital entities, in similar thinking as The Data Conservancy.

9 Application case snippets (incomplete/unreviewed)

These snippets are intended to provide some examples of the principles in action. They are of the form:

"If a digital resource is <PRINCIPLE>, as a <USER-ROLE> I should be able to <ACTION> in order to <BENEFIT>"

NOTE: These are an initial set of cases and will be further developed pending interaction across HMC Hubs and Projects.

Findable

1. If a digital resource is "Findable", as a general user I should be able to find all research data that match a given set/subset of metadata in order to restrict my search.
2. If a digital resource is "Findable", a user will be able to link metadata standards, like a common file format, such that the data is actionable by software using the metadata standard.
3. If a digital resource is "Findable", a user will be able to cite the resource using a globally persistent, unique identifier in publications.
4. If a digital resource is "Findable", a data scientist will be able to build specialised search indices based on metadata closely linked to the persistent identifier of resources in order to provide application-case driven search.
5. If a digital resource is "Findable", a data curator is able to identify relevant properties, e.g. data formats, in order to perform data curation tasks, e.g. data format conversion.
6. If a digital resource is "Findable", a metadata provider is able to assign tombstone information in order to mark the resource to be deleted.

Accessible

1. If a digital resource is “Accessible”, as a scientist I should be able to retrieve the metadata with a free protocol in order to access research cohorts.
2. If a digital resource is “Accessible”, as a general user I should be able to read or download the metadata in order to analyse its content.
3. To keep a digital resource “Accessible”, as a metadata repository provider I should guarantee easy export of metadata in order to preserve metadata beyond the lifespan of the repository.
4. Even if a digital resource is already “Accessible”, as a metadata provider and data producer (i.e. the rights holder of the data) I should be able to decide on which location the corresponding data is stored, e.g., to migrate to cheaper storage in a long run.
5. If a digital resource is “Accessible”, as a software developer I should be able to write plugins that allow loading data with standardised formats into appropriate tools.
6. If a digital resource is “Accessible”, as a general user I should be able to stage data (e.g. retrieve it from a repository) with standardised formats on local or remote machines in order to access them with appropriate tools.

Interoperable

1. If a digital resource is “Interoperable”, as a general user of the data I should be able to compare its metadata to those of other digital resources to accurately and collectively process the resources.
2. If a digital resource is “Interoperable”, as a scientist I should be able to technically use data on the same subject from different sources in order to compare them on a scientific basis.
3. If a digital resource is “Interoperable”, as a scientist I should be able to combine data from different domains in order to study cross-linked systems and interdependent phenomena and processes.

Reusable

1. If a digital resource is “Reusable”, as a scientist of the same field as the data creator I should be able to fully understand how it was created in order to continue doing science with it.
2. If a digital resource is “Reusable”, as a scientist of a different field as the data creator I should be able to sufficiently understand how it was created in order to continue doing science with it.
3. If a digital resource is “Reusable”, as a non-scientist I should be able to understand the quality and trustworthiness of the data in order to use it for my purpose with trust.
4. If a digital resource is “Reusable”, as a general user of the data I should be able to technically process the metadata in order to work with the data.

5. If a digital resource is “Reusable”, as a scientist of the same field or of a different field as the data creator I should be able to understand my legal rights regarding the use of the data in order to generate new results and publish them.
6. If a digital resource is “Reusable”, as a scientist of the same field or of a different field as the data creator I should be able to correctly cite the data or publications that describe them in order to publish own results based on the data.
7. If a digital resource is “Reusable”, as a general user of the data I should be able to understand the data semantically in order to interpret them and the results of calculations based on them.
8. If a digital resource is “Reusable”, as a scientist I should be able to semantically understand data on the same subject from different sources in order to compare them on a scientific basis.

References

Preamble:

[1] The FAIR Guiding Principles: https://www.force11.org/group/fairgroup/fairprinciples

[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.; The FAIR Guiding Principles for scientific data management and stewardship, Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[3] GO FAIR: https://www.go-fair.org/fair-principles/

[4] RDA (Research Data Alliance): https://www.rd-alliance.org/

[5] FAIR Data Maturity Model Working Group; FAIR Data Maturity Model - Specification and Guidelines, version 1.0 (2020). https://doi.org/10.15497/rda00050

[6] SNF (Swiss National Science Foundation); Explanations of the FAIR data principles: http://www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf

[7] Helmholtz Incubator platform Helmholtz.AI: https://www.helmholtz.ai/

[8] Helmholtz Incubator platform Helmholtz Imaging Platform (HIP): https://www.helmholtz-imaging.de/

Actions moving forward:

[9] W3C Working Group; An Overview of the PROV Family of Documents (2013): https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/

Definition:

[10] Wikipedia contributors; Communication protocol, Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Communication_protocol&oldid=1013721013 (accessed on the March 23, 2021).

[11] The World Wide Web Consortium (W3C), Standards: https://www.w3.org/standards/

Accessibility:

[12] Open Archives Initiative; Open Archives Initiative Protocol for Metadata Harvesting (OAI PMH), version 2.0 (2015): http://www.openarchives.org/OAI/openarchivesprotocol.html

[13] Open Archives Initiative; ResourceSync Framework Specification (2017): http://www.openarchives.org/rs/1.1/resourcesync

[14] DONA Foundation; Digital Object Interface Protocol Specification, version 2 (2018): https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf

Interoperability:

[15] Biodiversity Information Standards (TDWG); Darwin Core: https://dwc.tdwg.org/

[16] Unidata; Network Common Data Form (netCDF): https://www.unidata.ucar.edu/software/netcdf/

[17] Open Geospatial Consortium (OGS); OGS Standards: https://www.ogc.org/docs/is

[18] Ontobee: http://www.ontobee.org/
[19] European Bioinformatics Institute (EMBL-EBI), Ontology Lookup Service (OLS): https://www.ebi.ac.uk/ols/index
[20] United Nations Information Portal on Multilateral Environmental Agreements (InforMEA), Glossary: https://www.informea.org/en/terms/alphabetic
[21] Le Franc, Y., Parland-von Essen, J., Bonino, L. et al.; FAIRsFAIR - D2.2 FAIR Semantics: First recommendations, version 1.0 (2020). https://doi.org/10.5281/zenodo.5361930

Reusability:

[22] Global Indigenous Data Alliance (GIDA), CARE Principles for Indigenous Data Governance (CARE): https://www.gida-global.org/care
[23] Wikipedia contributors; Minimum information standard, Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Minimum_information_standard&oldid=1038415640 (accessed December 9, 2021).
[24] Licence selector: https://ufal.github.io/public-license-selector/
[25] Dublin Core Metadata Initiatives (DC); Properties of the term `license` (DC:terms: licence): https://www.dublincore.org/specifications/dublin-core/dcmi-terms/#terms-license