Project Title	Global cooperation on FAIR data policy and practice
Project Acronym	WorldFAIR
Grant Agreement No	101058393
Instrument	HORIZON-WIDERA-2021-ERA-01
Topic, type of action	HORIZON-WIDERA-2021-ERA-01-41 HORIZON Coordination and Support Actions
Start Date of Project	2022-06-01
Duration of Project	24 months
Project Website	http://worldfair-project.eu

D2.3: Cross-Domain Interoperability Framework (CDIF)

(Report Synthesising Recommendations for Disciplines and Cross-Disciplinary Research Areas)

Work Package	WP02 - Recommendations, Synthesis and FAIR Assessment
Lead Author (Org)	Arofan Gregory (CODATA)
Contributing Author(s) (Org)	Darren Bell (UKDA), Dan Brickley (Independent Expert), Pier Luigi Buttigieg (AWI, ODIS), Simon Cox (Independent Expert, OGC), Michelle Edwards (University of Guelph), Doug Fils (Ocean Experts), Luis Gerardo Gonzalez Morales (UN Stats), Pascal Heus (Postman), Simon Hodson (CODATA), Chifundo Kanjala (UNICEF),

	Yann Le Franc (eScience Factory), Lauren Maxwell (University of Heidelberg), Laura Molloy (CODATA), Steve Richard (Independent Expert), Flavio Rizzolo (Stats Can), Peter Winstanley (Semantic Arts), Lesley Wyborn (ANU).
Due Date	30.05.2024
Date	31.05.2024
Version	1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.11236871

Dissemination Level

- [X] PU: Public
- [ ] PP: Restricted to other programme participants (including the Commission)
- [ ] RE: Restricted to a group specified by the consortium (including the Commission)
- [ ] CO: Confidential, only for members of the consortium (including the Commission)

Versioning and contribution history

Version	Date	Authors	Notes
0.1	07.05.2024	Arofan Gregory, Simon Hodson, Luis Gonzalez-Morales, Flavio Rizzolo, Steve Richard, Doug Fils	Draft for internal review
0.2	29.05.2024	Arofan Gregory, Simon Hodson, Luis Gonzalez-Morales, Flavio Rizzolo, Steve Richard, Doug Fils, Pascal Heus, Lesley Wyborn	Final draft
1	29.05.2024	Arofan Gregory, Simon Hodson, Laura Molloy	Content finalised

Disclaimer

WorldFAIR has received funding from the European Commission's WIDERA coordination and support programme under the Grant Agreement no. 101058393. The content of this document does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of such content.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Abbreviations and Acronyms

API	Application Programming Interface
BPMN	Business Process Modelling and Notation
CC	Creative Commons
CDIF	Cross-Domain Interoperability Framework
CERIF	Common European Research Information Format
CESSDA	Consortium of European Social Science Data Archives
CODATA	Committee on Data of the International Science Council
CORDIS	Community Research and Development Information Service
CRS	Coordinate Reference System
CSV	Comma-Separated Values
CWL	Common Workflow Language
DCAT	Data Catalog Vocabulary
DDI	Data Documentation Initiative
DDI-CDI	Data Documentation Initiative - Cross-Domain Integration
DGG	Discrete Global Grids
DOI	Digital Object Identifier
DPV	Data Privacy Vocabulary
DRR	Disaster Risk Reduction
DwC	Darwin Core
EML	Ecological Metadata Language
ENVO	Environment Ontology
EOSC	European Open Science Cloud
ESS	European Social Survey
FAIR	Findable, Accessible, Interoperable, Reusable
FAO	Food and Agriculture Organization of the United Nations
FDOF	FAIR Digital Object Framework
FER	FAIR Enabling Resource

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FIP	FAIR Implementation Profile
FRBR	Functional Requirements for Bibliographic Records
GBIF	Global Biodiversity Information Facility
GFF	GO FAIR Foundation
HDF5	Hierarchical Data Format 5
HEIs	Higher Education Institutions
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
IGSN	International Generic Sample Numbers
ILO	International Labour Organisation
InChI	International Chemical Identifier
INSPIRE	Implementation Network for Sharing Population Information from Research Entities
IODE	Oceanographic Data and Information Exchange
IUPAC	International Union of Pure and Applied Chemistry
JSON-LD	JavaScript Object Notation for Linked Data
LLM	Large Language Model
NetCDF	network Common Data Form
NUTS	Nomenclature of Territorial Units for Statistics
OBO	Open Biological and Biomedical Ontologies
ODIS	Ocean Data Information System
ODRL	Open Digital Rights Language
OHDSI	Observational Health Data Sciences and Informatics
OGC	Open Geospatial Consortium
OMOP CDM	Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)
ORCID	Open Researcher and Contributor ID
OWL	Web Ontology Language
PARC	Partnership for Assessment of the Risks of Chemicals
PID	Persistent Identifier

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

PPI	Plant-Pollinator Interactions
PROV-O	PROV (Provenance) Ontology
QUDT	Quantities, Units, Dimensions and Types
RDA	Research Data Alliance
RDF	Resource Description Framework
REST	REpresentational State Transfer
ROR	Research Organization Registry
SALURBAL	Salud Urbana en América Latina
SDG	Sustainable Development Goal
SDMX	Statistical Data and Metadata Exchange
SDTH	Structured Data Transformation History
SDTL	Structured Data Transformation Language
SHACL	Shapes Constraint Language
SKOS	Simple Knowledge Organization System
SSSOM	Simple Standard for Sharing Ontology Mappings
TDWG	Biodiversity Information Standards (formerly the Taxonomic Databases Working Group)
TRE	Trusted Research Environment
UOM	Unit of Measurement
VTL	Validation and Transformation Language
W3C	World Wide Web Consortium
WHO	World Health Organisation
WKT	Well Known Text Representation of Geometry

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Executive summary

Increasingly, many important research questions demand a multi-disciplinary approach in which data and resources are used across domain and infrastructure boundaries. In such scenarios, domain-specific community standards fall short of the requirements for FAIR exchange of the critical metadata and other information needed. The Cross-Domain Interoperability Framework (CDIF) is designed to support FAIR implementation for these projects by establishing a 'lingua franca' for this information, based on existing standards and technologies to support interoperability, in both human- and machine-actionable fashion. CDIF is a set of implementation recommendations, based on profiles of common, domain-neutral metadata standards which are aligned to work together to support core functions required by FAIR.

The idea for CDIF first emerged from workshops and discussions at conferences prior to the WorldFAIR project, beginning in 2018. The WorldFAIR project provided an opportunity to advance that vision, through a set of 11 case studies across many domains, allowing the needs and practices around FAIR within such domains to be summarised in the form of FAIR Implementation Profiles (FIPs). Based on the FIPs and focused meetings, the requirements for CDIF were established. A group of 30 invited experts from different FAIR initiatives and standards bodies made up a Working Group and an Advisory Group to synthesise the findings from WorldFAIR and to produce the current CDIF draft. Work on CDIF will continue after the WorldFAIR project and CODATA will maintain the WG and AG.

This report presents a core set of five CDIF profiles, which address the most important functions for cross-domain FAIR implementation.

1. **Discovery** (discovery of data and metadata resources)
2. **Data access** (specifically, machine-actionable descriptions of access conditions and permitted use)
3. **Controlled vocabularies** (good practices for the publication of controlled vocabularies and semantic artefacts)
4. **Data integration** (description of the structural and semantic aspects of data to make it integration-ready)
5. **Universals** (the description of 'universal' elements, time, geography, and units of measurement).

Each of these profiles is supported by specific recommendations, including the set of metadata fields in specific standards to use, and the method of implementation to be employed for machine-level interoperability.

A further set of topics is examined, establishing the priorities for further work. These include:

1. **Provenance** (the description of provenance and processing)
2. **Context** (the description of 'context' in the form of dependencies between fields within the data and a description of the research setting)
3. **Perspectives on AI** (discussing the impacts of AI and the role that metadata can play)
4. **Packaging** (the creation of archival and dissemination packages)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

5. **Additional Data Formats** (support for some of the data formats not fully supported in the initial release, such as NetCDF[1], Parquet[2], and HDF5[3]).

In each of these topics, current discussions are documented, and considerations for further work are provided.

CDIF is designed to leverage the work of other FAIR initiatives such as FAIR-Impact and the work in EOSC. It is designed to be implementable with existing tools, standards, and technologies but, as a set of recommended practices, must be maintained whilst the state of play around FAIR implementation develops and evolution occurs in the technology sphere. CDIF leverages methodologies such as FIPs from the GO FAIR Foundation[4]. Importantly, it aligns with efforts such as the EOSC Interoperability Framework[5], and developments such as Signposting[6] and reference implementation of the FAIR Digital Object Framework[7]. Work on semantic mapping and in some other areas is informed by on-going developments in other fora such as RDA[8]. CDIF is designed to enable the practical implementation of FAIR by supporting these frameworks and approaches in cross-domain scenarios.

In any given domain, the standards used should be to a considerable extent mappable to and from their corresponding CDIF profiles, reducing the volume of mappings needed to interoperate effectively for core FAIR functions in multi-disciplinary scenarios. Broadly speaking, FAIR demands an increase in the metadata provided by the disseminators of data, especially if we are to automate resource-intensive data integration tasks, which today are largely manual. In a cross-domain scenario, the sheer number of mappings needed is not supportable. **CDIF provides a solution by changing a many-to-many dynamic into a many-to-one dynamic.**

CDIF is not intended to replace existing community standards, but to supplement them for communication across domain and infrastructure boundaries. It does not aim to replace the specific models needed within different domains, but it does aim to establish a foundation of common metadata which can support a core set of FAIR functionality. Real-world examples of large-scale standards-based exchange networks, such as the Statistical Data and Metadata Exchange[9] (SDMX) and the Ocean InfoHub[10] (ODIS) have been used as inspiration for the overall approach, to ensure its feasibility for practical implementation. This draft includes links to early prototypes for such data as the Sustainable Development Goal Indicators and some of their source data, showing how the mining of the native standard descriptions of the source, to produce its equivalent in CDIF, can support disaggregation and integration of that data with other sources.

[1] https://en.wikipedia.org/wiki/NetCDF
[2] https://parquet.apache.org/
[3] https://www.hdfgroup.org/solutions/hdf5/
[4] https://www.go-fair.org/how-to-go-fair/fair-implementation-profile/
[5] https://op.europa.eu/en/publication-detail/-/publication/d787ea54-6a87-11eb-aeb5-01aa75ed71a1/language-en
[6] https://signposting.org/FAIR/
[7] https://fairdo.org/
[8] https://www.rd-alliance.org/
[9] https://sdmx.org/
[10] https://oceaninfohub.org/odis/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

This document establishes the core set of profiles for the highest priorities in cross-domain FAIR implementation, at a practical level. Over time, this core set of functions will grow to form the basis of broader support in this critical area of FAIR exchange and use.

Summary of CDIF profiles and recommendations

General: CDIF metadata should be embedded in landing pages or linked stand-alone files, encoded in JSON-LD. The supported profiles will be indicated as part of the metadata.

Discovery profile: This profile recommends the use of a set of key Schema.org fields for describing static data sets and queryable data sources, with the DCAT[11] equivalent recognised as an acceptable alternative.

Access profile: This profile recommends that ODRL[12] Actions and Entities be used to describe policies and conditions for the use of data. At this time, the utility of this approach is limited by the lack of shared vocabularies for conditions of use, user qualifications, legal constraints, and similar important items. ODRL is thus limited to describing policies in terms of the disseminating institution, but provides a basis for expansion in future when the needed vocabularies are developed.

Controlled vocabularies profile: This profile recommends the use of SKOS[13] for describing controlled vocabularies, understood to mean any terminological resource. The use of OWL[14] as a linked extension to what is presented in SKOS is also recommended, as is the use of XKOS[15] for formal statistical classifications.

Data description profile: This profile recommends the use of DDI-CDI[16] to provide a granular description of the structure of data sets, and how the logical content of those datasets relates to their physical encoding. Text-based data is supported (CSV and other delimited formats, fixed-width ASCII, etc.), with the intention of expanding support for other types of data in future. The recommendations cover description of individual data sets to make them 'integration-ready'.

Universals profile: This section recommends the information which should be provided when describing time, geography, and units of measurement in other metadata sets. Some standards for this purpose are recommended in each area.

[11] https://www.w3.org/TR/vocab-dcat-3/
[12] https://www.w3.org/TR/odrl-model/
[13] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[14] https://www.w3.org/TR/owl2-overview/
[15] https://ddialliance.org/Specification/RDF/XKOS
[16] https://ddialliance.org/Specification/ddi-cdi

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Table of contents

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

0. Definitions

Definitions in this section follow the conventions outlined in ISO704[17]. Definitions are derived through interpretation of Fair Data Object Framework documents[18] and Kahn and Wilensky (2006)[19], unless otherwise noted.

Digital Object — packaged, identifiable sequence of digital bits that carries some information. A digital object has exactly one digital representation. Note that 'Digital Object' is capitalised in this document to emphasise that it is being used in this specific sense.

FAIR — *F*indable, *A*ccessible, *I*nteroperable, and *R*eusable. [20]

FAIR Digital Object — **Digital Object** in the context of a system of policies and infrastructure to support realisation of the FAIR principles.

identifier (CDIF) — an association between a sequence of bits and a unique resource of interest[21]. Typically represented as a sequence of characters. **Note**: identifiers lose value if there is no clear, unambiguous specification of what they identify.

key-value — a system in which data values are individually paired with identifiers ("keys"), without reliance on the containing structure or other information to disambiguate them. Typical of some "big data" technologies, data lakes, etc.

persistent identifier — **identifier** that has the intention that its binding is to the precise same resource for the lifetime of systems using the identifier.

PID — **persistent, registered identifier** that can be resolved to obtain a **PID kernel record**.

PID kernel record — **Digital Object** that provides documentation for the source of the PID, expected lifetime, linkage to the resource it represents, and other attributes specified in PR-KernelAtributues-2.0[22].

PID profile — **Digital Object** that provides 1) the definition of the protocol for obtaining the **PID kernel record** given the **PID**, and 2) the definition of the schema for the content of the PID kernel record.

PID registration — binding between a **PID** and a **PID kernel record** that provides documentation for the source of the PID, expected lifetime, and linkage to the resource it represents. To be useful, there must be a known community practice for obtaining the PID kernel record given the PID.

population — universe in which the individuals share time and geography. Typically the individuals are located in the same temporal and spatial extent. A pan-European survey administered in France and Greece might have a distinct population for each country.[23]

registered digital object — **Digital Object** that has been placed in an accessible digital storage location, and assigned a **registered and resolvable identifier** that can be resolved to access the object.

---

[17] https://edisciplinas.usp.br/pluginfile.php/312607/mod_resource/content/1/ISO%20704.pdf
[18] PR-PIDProfileAttributes-2.0-20220608 (https://docs.google.com/document/d/1c2mZziq5pIPmLxMHLcYqlWrjYsc2ezGMXvp0E46iljo); Bonino, L., Guizzardi, G., Sales, T., (2022) 'FAIR Digital Object Framework Documentation', https://fairdigitalobjectframework.org/; PR-KernelAtributues-2.0 (https://docs.google.com/document/d/1OF49wTNVuv-6OXlNerhBTqVtHyc7jutTaUHjn6BZCs0)
[19] https://doi.org/10.1007/s00799-005-0128-x
[20] See https://www.go-fair.org/fair-principles/
[21] https://www.ietf.org/archive/id/draft-kunze-ark-37.html#name-definition-of-identifier-2
[22] https://docs.google.com/document/d/1OF49wTNVuv-6OXlNerhBTqVtHyc7jutTaUHjn6BZCs0
[23] Adapted from the draft of the upcoming revision of the DDI Glossary. https://ddialliance.org/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

registered identifier — **identifier** that has been recorded in an accessible information system with its binding to the resource it identifies.

resolvable identifier — **identifier** that has a known protocol to access the resource to which it is bound. In current standard practice this is achieved using the World Wide Web and its associated HTTP(s) protocol.

semantic artefact — machine-actionable, human-readable representation of an abstract, simplified view of the world[24]

unit — individual member of a group that is being measured. A distinct member of a **universe**. Technically, an individual whose properties correspond to the characteristics of a **unit type**. Properties distinguish **units**; characteristics distinguish **unit types.**[25]

unit type — class of **units** defined by essential characteristics. Examples: person, household, business establishment. A kind of entity based on a set of characteristics. The characteristics are specific to the study context. Ideally a category of individuals in a non-overlapping classification. [26]

universe — class of individuals that share a **unit type** and typically have other characteristics in common, exclusive of time and geography. Given a **unit type**, a **universe** is the domain of those '**units**' that is observed. For example a **unit type** might be 'humans' and a **universe** might be 'humans who are nurses'. A **population** is a **universe** at a given place and time. Some definitions of **universe** are more or less precise about the characteristics that differentiate between 'universe' and 'population'.[27]

variable — mapping to a value domain from a set of **units** corresponding to a single **unit type**. A mapping is an association. This usage does not correspond exactly with other common uses (e.g. variable and attribute (research)), but is specific to data description. If the **unit** is a person, their gender would be a characteristic. The values of these characteristics are also concepts, e.g. 'male'. The variable here is 'gender', the value 'male' is a category. Values that a variable assigns are properties of the **units** being measured.[28]

[24] Based on FAIRsFAIR D2.8 FAIR Semantics Recommendations, Third Iteration, H2020-INFRAEOSC-2018-4, https://zenodo.org/doi/10.5281/zenodo.6276576
[25] Adapted from the draft of the upcoming revision of the DDI Glossary. https://ddialliance.org/
[26] Adapted from the draft of the upcoming revision of the DDI Glossary. https://ddialliance.org/
[27] Adapted from the draft of the upcoming revision of the DDI Glossary. https://ddialliance.org/
[28] Adapted from the draft of the upcoming revision of the DDI Glossary. https://ddialliance.org/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

1. Introduction

The Cross-Domain Interoperability Framework (CDIF) is a set of guidelines and practices for using domain-agnostic standards to support the interoperability and reusability of FAIR data, especially across domain and institutional boundaries. It has been developed in response to the need for agreements on the use of standards in FAIR implementations, to support all areas of research but in particular those grand challenge research questions that are necessarily interdisciplinary and cross-domain in nature, such as climate change, natural disasters, infectious disease, and so on. CDIF is intended for implementers of systems which support this type of FAIR use and exchange, and will be of use particularly to a technical audience.

The development of CDIF was informed by the work of the eleven WorldFAIR Case Studies and their FAIR Implementation Profiles (FIPs): the CODATA team held intensive meetings with each of the case studies to understand practices and challenges in data and metadata. A number of CDIF modules were advanced through two Dagstuhl workshops (organised in collaboration with the DDI Alliance[29]), held during the WorldFAIR project and at which all case studies participated, alongside other experts.

Additionally, thirty invited experts participated in drafting the CDIF guidelines, including members of many related FAIR initiatives and standards bodies. Therefore, the CDIF guidelines draw on significant expertise both within and outside the WorldFAIR project.

Just as CDIF is designed to complement and add detail to the EOSC Interoperability Framework[30], as well as work by other EOSC-related projects, it is also informed by a number of global developments represented in WorldFAIR. CDIF is designed to build on the work of such groups, complementing their efforts and making them practical for implementers in Europe and globally.

Work on CDIF will continue beyond the life of the WorldFAIR project, through a series of follow-up projects (case studies and implementation pilots) coordinated as WorldFAIR Plus (WorldFAIR+). The CDIF guidelines are only meaningful in relation to developments in the realm of standards and technology, and must be maintained accordingly[31]. The CDIF Working Group and Advisory Group will be maintained by CODATA, as will the strong relations with standards bodies, authoritative organisations and international, regional and national infrastructures who are the key stakeholders.

1.1. How to read this document

In what follows, Sections 2-5 (2. Overview; 3. Why CDIF matters; 4. How was CDIF developed?; 5. What is the structure of CDIF?) provides an introduction to CDIF, its origins and purpose. These sections are intended to inform the wider data stewardship community. We also intend these sections to be of utility to policy makers (to understand why CDIF matters and what its benefits can be) and managers of data infrastructures (who would oversee the implementation of CDIF to support cross-domain functions).

[29] https://ddialliance.org/ - see “Cross-Domain Approaches”.
[30] https://op.europa.eu/en/publication-detail/-/publication/d787ea54-6a87-11eb-aeb5-01aa75ed71a1/language-en/format-PDF/
[31] https://worldfair-project.eu/cross-domain-interoperability-framework/

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Following a discussion of the design approach taken and good practice for the technical expression of metadata, the **core set of five CDIF profiles** are presented in Section 6. These address the most important functions for cross-domain FAIR implementation.

1. **Discovery** (discovery of data and metadata resources)
2. **Access** (specifically, machine-actionable descriptions of access conditions and permitted use)
3. **Controlled vocabularies** (good practices for the publication of controlled vocabularies and semantic artefacts)
4. **Data integration** (description of the structural and semantic aspects of data to make it integration-ready)
5. **Universals** (the description of 'universal' elements, time, geography, and units of measurement)

The chapters describing these profiles include implementation guidelines and code examples and are primarily intended for implementers, including technical staff at data infrastructures. As this is the case, each chapter concludes with a targeted list of resources (including reference documents and other material) to support implementers in addressing the profile.

Section 7 briefly discusses topics for future work and is intended to inform a more general data stewardship audience of the directions work on CDIF will follow.

Finally, the appendices provide additional implementation guidance and some specific examples. They are again intended for implementers and technical staff in data infrastructures.

2. Overview: what is the Cross-Domain Interoperability Framework (CDIF)?

2.1. What is the aim of the CDIF?

CDIF is designed to address one of the questions which arises during implementation of systems in accordance with the FAIR Data Principles[32]: what standards should be used, and how should they be implemented? The FAIR principles require that data and metadata be described according to common standards and made available through common protocols and mechanisms. As principles, appropriately, they do not specify *which* standards and protocols, leaving the answers to these questions up to the implementers. Within the context of a specific domain or infrastructure, established practice can provide some guidance: implementers can use the standards and protocols which are common within that domain or infrastructure.

2.2. The challenge of cross-disciplinary FAIR data

In a scenario where FAIR resources are intended for use across domain and infrastructure boundaries, however, this approach breaks down: do we expect systems in every domain to support the standards and

[32] 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

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

protocols of every other domain whose resources might be of interest? When we consider the number of different functions which need to be supported for the discovery, access, integration, and use of data, and the number of standards and protocols these functions require, this becomes an overwhelming task.

CDIF attempts to provide recommendations for implementers to help resolve this issue: if we can agree on a set of non-domain-specific standards and their implementations for scenarios where FAIR exchanges are taking place across domain and infrastructure boundaries, then the scope of the problem can be materially reduced. Large-scale networks rely on standard information. (Consider the use of HTML for the Web: there is a single mark-up language for the display of basic information in a browser.) While there is no authority which can dictate how the FAIR principles are to be implemented, it is possible to recommend a single approach which will promote convergence on a set of standards, at least for typical cases. While adoption of such recommendations is of necessity voluntary, the utility of a common approach will serve as the motivator.

2.3. Implementing FAIR for cross-domain data

There is a wealth of research on how FAIR can best be implemented, and investigations into this subject show no sign of abating. CDIF is not a research project, nor does it intend to outline the *best* possible solution to the challenges of FAIR implementation. Instead, it is first and foremost a practical exercise: how can we find a set of common standards and technology approaches which will enable us to implement FAIR for cross-domain scenarios using things which exist today? Standards are only useful if they are agreed and implemented, so CDIF tries to advocate those standards and technologies which are already in common use to the greatest extent possible. In some cases, there are gaps in current practice which require additional information and standards, but these are minority cases. Web standards already exist for supporting most of the needed functions, but they are not always used in ways which are interoperable, requiring a common approach for at least some core aspects. The CDIF recommendations are based on practical considerations: we must agree on a body of practice, and whether it is the best possible solution is a secondary concern. More important is that it be implemented widely, so that FAIR exchange can become as easy as possible. Once laid, such a foundation can be perfected. A long-running dispute over the best approach to FAIR implementation seems unlikely to provide a practical solution for immediate use.

2.4. Who can use the CDIF?

CDIF is aimed primarily at data infrastructures, i.e. those organisations which develop, maintain, and disseminate FAIR resources for reuse, often as centralised points of access within their communities or area(s) of interest. While data stewardship by research organisations is an important element of FAIR, not all research organisations perform this dissemination, instead relying on data archives or other dedicated repositories. FAIR reuse is most effective when authoritative producers, or those acting on their behalf, provide their data and metadata to others for reuse, so the authoritative versions of such resources are the ones which get reused. Such organisations are often motivated to be the point of dissemination, as it is their mission, and they bear the responsibility, both legal and reputational, for those resources. CDIF is a tool which they can use to better support this mission.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

2.5. A set of common functions

To this end, CDIF identifies a set of common functions which are needed to implement FAIR: describing data for discovery and assessment, supporting access to data by describing the licensing and conditions of use, laying out the structure and semantics of data to support automated integration, and providing information regarding the provenance of data and resources. In each area, practical recommendations are made regarding what standard or standards should be used, and how they should be implemented. In those areas where there is no clear practice which can be recommended, or which require further investigation, this is noted. For the provision of basic discovery metadata, the description of licence conditions, the publication of controlled vocabularies, and the description of data to make it 'integration ready', specific steps are described which can be used to guide immediate implementation.

2.6. A set of profiles for different users

The CDIF recommendations are presented as a set of related profiles of domain-agnostic standards, with a specific implementation approach. Web technologies provide the basis for this approach. While many of the standards and vocabularies require the use of RDF, it is not a technology which is commonly used in every domain. The solution to this is to advocate the use of JSON-LD, which allows the expression of RDF vocabularies in the common JSON syntax.

In each profile, a minimum set of required fields are indicated, to support common cases. Other optional fields are suggested, and the path forward toward support of more complex scenarios is indicated. While FAIR implementation is demanding, it is hoped that by identifying a common core of metadata, it can be made as cost-effective as possible.

It is intended that users will adopt only those profiles which are useful to them. There is no requirement for the adoption of unneeded profiles, except where these dependencies exist for practical reasons. Thus, it is possible to describe data to make it 'integration ready' at a detailed level, but not to support profiles for data discovery or access, to give but one example. CDIF profiles are intended to be a toolkit for implementation, with the needed functions being addressed in any specific setting according to implementer priorities.

2.7. Sustaining the CDIF

Future considerations are presented: CDIF is intended to be developed further, and maintained over time. To this end, the work is guided by a set of explicit design principles, as it is recognized that technology and the available standards will change over time, and sometimes rapidly. There are many different initiatives looking at FAIR implementation, and CDIF will be maintained in alignment with these efforts as additional practice emerges. Developments in artificial intelligence cannot be ignored, although there is as yet no great clarity regarding their impact.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

3. Why the CDIF matters

The basic vision of FAIR is a simple one: research data and related resources should be easy to find, access, combine, and use. The implications of this vision are profound, in terms of better, more efficient and effective research. This idea applies especially to research which is multidisciplinary and which cuts across domain and infrastructure boundaries. Many of the biggest challenges we face today as societies demand that research realises this vision of efficient, cross-domain sharing of resources.

Yet the practical reality lags far behind the vision. Despite the promises of modern information technology, our ability to easily find and share research data is still very limited. This is not a problem of vision, or a problem of technology: we know what kind of research infrastructure we want to build, and we have the tools to build it. What is lacking is more prosaic: in practice, the needed resources are not always provided to researchers in a way which allows them to be easily reused. Despite a large and growing interest in the FAIR Principles, the reality has not kept pace. The barriers to progress stem not only from the culture around technology adoption, but also from organisational and legal challenges to collaboration.

Practical FAIR implementation requires a high level of standardisation in how data and metadata resources are exchanged between providers and users. While standards are increasingly being adopted in different domains, these standards are often specific to the domains which use them. This is FAIR, but not sufficient to support cross-domain use.

In many different examples, we have seen that research questions which seem as if they should be easy to answer present the researcher with barriers. A good example of this comes from the recent EOSC Future's 'Climate Neutral and Smart Cities' project: can social researchers with data about attitudes toward climate change compare this with the climate conditions being experienced by the respondents? All needed data is publicly available, and most of it documented in standard forms. Despite this, the effort required to produce the needed integrated data was unexpectedly large[33].

If we want researchers (or more accurately research teams and projects) to be able to locate and find data from different sources, and to combine it without requiring hours or days of preparatory work manually massaging the data into an integrated form, we need to provide detailed and rich metadata in standards which are meaningful to the user, and especially the user's systems. For FAIR implementation, machine-actionability is a key enabler - without automation, the problems of scale become overwhelming.

CDIF proposes a set of recommendations aimed at exactly this challenge: can we coordinate our use of *existing* standards and technologies to make FAIR data use easy in practice? If this problem can be solved, then the utility to the researcher will be potentially huge: time spent 'data wrangling', and the lack of useful standards comes at a tremendous cost in terms not only of researcher effort and resources, but also negatively impacts the quality of research. This in turn has an effect on those who benefit from the use of research, such as policy makers. The opportunity costs in 2018 for Europe alone were estimated at 10.2

[33] See https://www.europeansocialsurvey.org/esslabs and especially the 'Reuse and Proproducibility' paper: Gregory A, Wackerow J, Orten H (2023) Reuse and Reproducibility: Describing Cross-Domain Research Data in the Science Project Climate Neutral and Smart Cities. ARPHA Preprints. https://doi.org/10.3897/arphapreprints.e115047

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

billion EUR per year. The factors at play here are described in a report from the European Commission on the 'Cost of Not Having FAIR Research Data', which also estimated that around 80% of research team or research project effort went into pre-analysis activities of finding, preparing, and cleaning data.[34]

The advent of large language models (LLMs) and other AI technologies both promises and threatens to have an amplifying effect: good practice around FAIR data and resources can help protect us from the negative impacts of these technologies, and help us to realise their promise in a powerful way, although this is currently difficult to quantify.

From a more positive perspective, the possibilities of having a FAIR-enabled cross-domain research infrastructure are impressive. Some existing examples show us how standards-based, scalable networks of data providers can support research. The SDMX infrastructure employed by the UN system for the SDG Indicator data is one example.[35] Another is the Ocean Data Information System (ODIS)[36] (as is described more fully in the work of Work Package 11 on Ocean Science[37]). CDIF will combine with infrastructure efforts such as open science clouds to help make real the promise of such efforts: data which is immediately findable, accessible, and usable in accurate, responsible, and efficient ways.

4. How was the CDIF developed?

The idea for CDIF came initially from a series of workshops held at Schloss Dagstuhl in Wadern, Germany, starting in 2018. The focus was on the alignment of standards from different domains, and the events were co-hosted by the Data Documentation Initiative (DDI) Alliance, building on a long-running series of Dagstuhl events. Each year subsequently, further discussions were held[38].

Participants in the workshops included standards developers and implementers from a wide range of backgrounds, and a general effort was made to keep the work grounded by including participants from a range of scientific domains, as well as computer scientists and standards and infrastructure experts. While sometimes exploratory or theoretical, use cases from the real world were used as a basis for the work.

As the subject of how standards could be used together was explored further, it became clear that substantial interoperability across domain boundaries would be possible if some basic agreements could be forged on the content and expression of key metadata.

These discussions provided input to the planning for CODATA's 'Making Data Work for Cross-Domain Grand Challenges' programme[39], in which the ideas for a standards-based interoperability framework became more

[34] European Commission, Directorate-General for Research and Innovation, Cost-benefit analysis for FAIR research data – Cost of not having FAIR research data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/02999, p.4, p.13.
[35] For a brief overview, see https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.32/2020/mtg1/W_4_1_ENG_SDMX-for-SDGs.FInal.pdf
[36] See https://oceaninfohub.org/odis/
[37] https://zenodo.org/records/7682399
[38] A listing of the workshops can be found at https://codata.org/initiatives/decadal-programme2/dagstuhl-workshops/
[39] Sponsored by the International Science Council, as part of the ISC's Action Plan https://council.science/actionplan/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

fully formalised, in line with the recommendations from the EU's 'Turning FAIR into Reality' report[40]. While similar work such as the EOSC Interoperability Framework was ongoing, none of this work specifically addressed the field-level alignments across metadata standards which could produce machine-to-machine interoperability in a cross-domain scenario.

With the WorldFAIR project, it became possible to access a range of domain expertise around FAIR, and to ground the work on CDIF more broadly. Each of the WorldFAIR Case Studies completed an initial FAIR Implementation Profile, and this served as input to provide an overview of what standards were currently in use and which were being considered. The FIPs also provided a view into the degree of standards adoption across different communities. Subsequent meetings were held with each group to look at how well the domain-neutral standards being considered for inclusion in CDIF served to describe FAIR data and resources from those communities. These inputs proved to be valuable. To cite two examples:

- In Work Package 07 on Population Health, Schema.org was used extensively to describe the context of an experiment for the purposes of exchange, in line with work done by the Observational Health Data Sciences and Informatics (OHDSI) group (which provided the Observational Medical Outcomes Partnership (OMOP) Common Data Model used as a standard data description in that community). As in many other examples, the implementation of Schema.org relied on JSON-LD.
- The Work Package 11 Oceans Science and Sustainable Development Case Study provided a view of the way a large-scale knowledge graph could be developed on the basis of a common core of metadata fields, again using Schema.org embedded in landing pages as JSON-LD. More information on WP11's relation to CDIF is available in WorldFAIR D11.2[41].

Synergies such as these emerged from the work, and helped to guide the development of the CDIF recommendations. Other connections to the Case Studies may be less evident, because the CDIF functionality is truly broad in scope, and attempts to provide a solution for all domains, and not just those represented in the project. This is reflected in the use of standards and technologies which are seen as widely used across all domains. The issues being addressed by CDIF apply to all domains, and so may not derive directly from specific statements reflected in any individual Case Study.

The work on CDIF was primarily conducted by a group of 30 invited experts, divided into a Working Group and an Advisory Group. Representatives from many different FAIR initiatives, implementations and standards bodies were included, many of them having attended the Dagstuhl workshops in preceding years. The Working Group met every two weeks over the duration of the project, discussing and drafting specific recommendations which were then passed to the Advisory Group for review. Public review of initial drafts was also informally conducted in some cases.

An effort was made to consider additional successful examples in related fields. While there is no existing cross-domain network for FAIR data today, there are initiatives from the scientific world and the world of

[40] European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality – Final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524
[41] https://zenodo.org/records/7682399

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

official statistics which proved valuable in showing which approaches could be employed. Among these was the Statistical Data and Metadata eXchange (SDMX) Initiative[42], which provides a standards-based framework for the exchange of statistical data and metadata among national and international statistical agencies and central banks in a broad global network. Significantly, SDMX standards are employed in the collection and publication of the Sustainable Development Goals (SDG) Indicators[43]. The Helmholtz Metadata Collaboration (HMC) likewise served as an example of how a large number of disparate research centres in Germany could exchange data and metadata in a practical fashion, leveraging the ODIS approach. Several such examples were considered, with an eye toward identifying practical implementation approaches.

The criteria for selecting standards for cross-domain use include:

1. they should be domain-agnostic — that is, not based on assumptions specific to a particular domain;
2. they must be open;
3. they should be already in use for the intended purpose, to the greatest extent possible;
4. they should be adoptable — that is, there is a community which can support them, provide tools and knowledge, etc.; and,
5. they should support anticipated functions in the future, at least to a reasonable extent.

These criteria limit the number of candidates to a relatively narrow field, as much of the work on standardisation has been grounded in specific domains. Specifications intended for use on the Web are prominent, as these frequently meet the criteria given above.

CDIF has been developed with a knowledge that there cannot be — and that there likely never will be — a single set of standards which will be useful for all FAIR exchanges in all situations and across all domains. Despite this, it is clear that a much higher degree of convergence is possible. Within domains, standards-based approaches have proven to be successful in many different areas. The idea that the FAIR vision can be practically realised without a similar movement toward convergence, even if imperfect, does not seem plausible. CDIF attempts to start an ongoing process of defining what such a convergence can look like, based as much as possible in approaches which have been shown to work, and using standards and technologies which are fit for the purposes of cross-domain exchange.

5. What is the structure of the CDIF?

The organisation of the CDIF recommendations follows a basic functional breakdown which stems from the need to support certain exchanges of information as dictated by the FAIR Principles. ‘Findable’ requires that we be able to make our resources searchable, and enable them to be catalogued. ‘Accessible’ requires that we can retrieve or link to resources in a technical sense, but also that we can understand what conditions are in the case that access is limited. ‘Interoperable’ means that we can load resources into our systems for processing once acquired, and operate on them in a meaningful way. ‘Reusable’ means that we have enough information to understand the data and the uses to which it can legally be put.

[42] https://sdmx.org/

[43] https://unstats.un.org/sdgs/indicators/indicators-list/

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

There is not a point-for-point alignment between the FAIR Principles and the functional requirements which support them: the information needed to support one principle may be required also to support another, and there may be no clear distinction made between these sets of information. CDIF organises FAIR 'functions' in a fashion which maps to the principles, but more closely follows the interactions between systems needed for implementation and the information they require:

1. Discovery (F)
2. Access (A)
3. Publication of controlled vocabularies and mappings (F, I, R)
4. Description of data for integration purposes (F, A, I, R)
5. Universal metadata: time, geography, and units of measure (F, I, R)
6. Provenance and process description (I, R)
7. Contextual information: dependencies within the data (I, R).

This organisation is not strictly functional, but attempts to describe the information needed for each major function (Discovery, Access, Integration) as well as to address some of the common information needs (Controlled vocabularies and mappings, Universals). Not all topics are covered in equal depth at this time: the first five are supported more thoroughly in this version of the guidelines than the other topics, which will receive more attention in future. This reflects the current state of play within the FAIR community, and the perceived relative importance of these functions based on current and planned implementations.

These areas were identified through examination of FAIR implementations in many domains, and are driven by the relative maturity of the standards and practice within communities engaging proactively with FAIR. Discovery is perhaps the most common subject of FAIR implementation, as it is both less demanding in terms of metadata (and therefore resources) as well as being logically a first step: if you can't find it, you can't use it! Access to open data is in some senses a 'solved' problem, so the attention of the FAIR community is turning towards the need to better support access to controlled data. Currently, support for providing access to controlled data is often strictly manual, presenting a practical bottleneck for reuse. While we are early in our development of standards and systems for automating access to controlled data, there are some initial steps which can be easily taken. Data interoperability and reuse have been receiving an increasing amount of attention in many domains: these are arguably the most metadata-intensive aspects of FAIR, but they also hold a huge potential in terms of efficiency gains: if we can ease the problems of integration and harmonisation ('data wrangling') through automation, the potential resource savings are large. Data integration necessarily raises the question of how semantics are exposed and mapped. These topics provide the focus of the current document and the first five CDIF profiles.

We also discuss topics for future work. While it is clear that the description of data provenance and contextualising information is critical, these topics remain more difficult to approach. Our initial explorations have identified that there is a wide variety of practice, in some cases competing standards, and a lack of consensus on the necessary scope. Therefore, some directions are suggested for consideration and future work.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Some topics are also discussed which are not directly functional, but which deserve mention. Foremost among these is how CDIF relates to developments around large language models (LLMs). Also, some aspects of FAIR implementation beg larger questions regarding how organisations understand their own roles vis-a-vis the FAIR resources they provide: how do we understand what constitutes a repository? What information does such an organisation need to provide as regards the resources it disseminates? Additionally, there are some data formats commonly used with very large data sets (NetCDF, Parquet, HDF5, etc.) which are not commonly described in RDF or other Web languages, and this subject represents an important topic for the future.

The initial five core CDIF profiles represent a strong starting point for cross-domain FAIR implementation. Our intention is that CDIF will be improved and extended as more FAIR implementation occurs.

6. CDIF Core Profiles

This section describes the set of five core CDIF profiles which we recommend to support different FAIR functions. At this point, many domains are exploring how FAIR can be implemented, and what standards and technologies are suitable. To assist this endeavour, **the core set of five CDIF profiles address the most important functions for cross-domain FAIR implementation.**

1. **Discovery** (discovery of data and metadata resources)
2. **Access** (specifically, machine-actionable descriptions of access conditions and permitted use)
3. **Controlled vocabularies** (good practices for the publication of controlled vocabularies and semantic artefacts)
4. **Data integration** (description of the structural and semantic aspects of data to make it integration-ready)
5. **Universals** (the description of 'universal' elements, time, geography, and units of measurement)

The diagram below shows what is covered in this initial draft of the CDIF recommendations.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

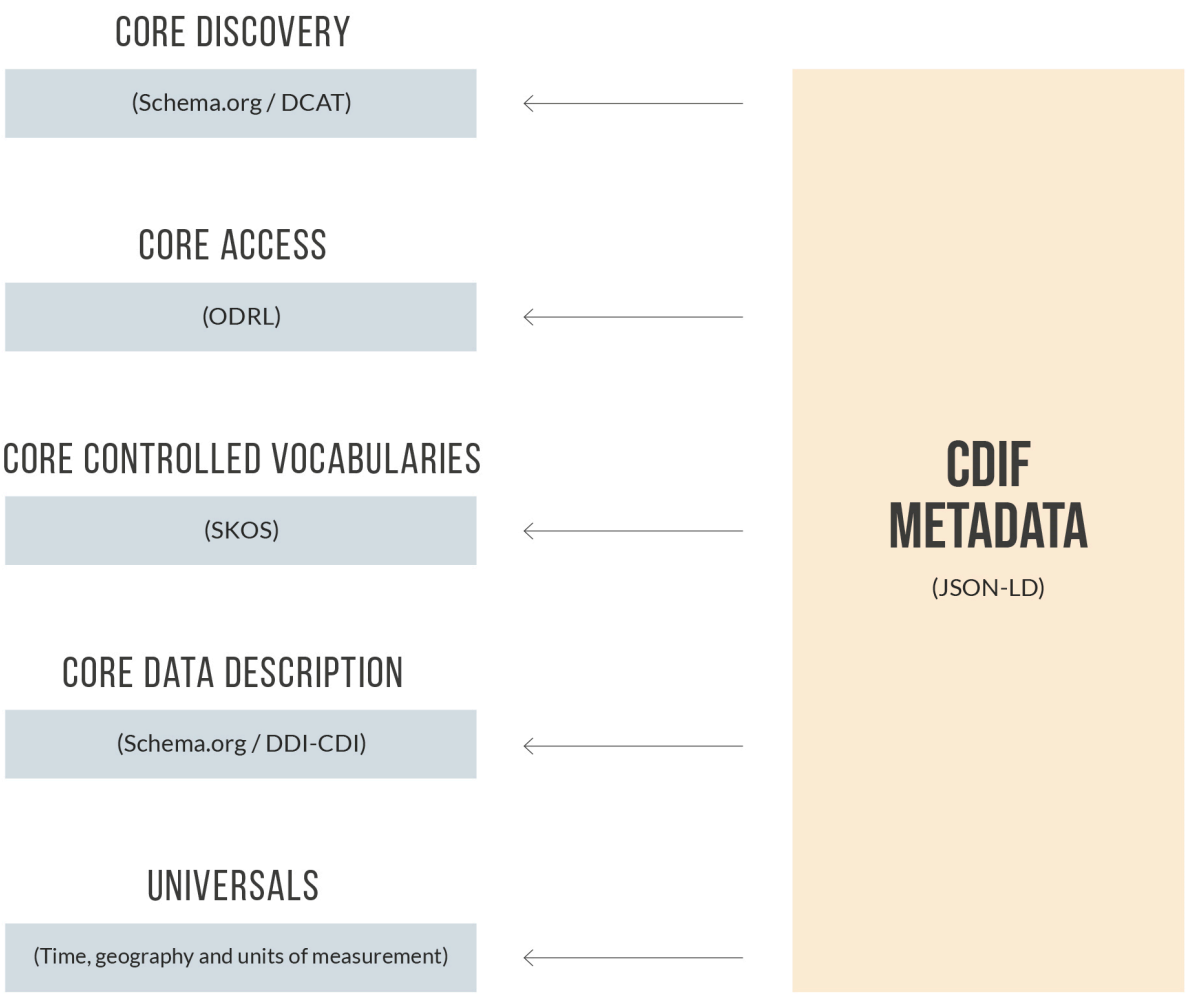

Figure 1a. CDIF Core Profiles.

Each of the listed profiles on the left-hand side of the diagram represents a set of metadata used to support a specific function. The exception to this is the 'Universals', which specifies the conventions for addressing specific types of metadata which can occur in any of the other profiles. The sets of recommended metadata are encoded as JSON-LD and expressed in a standard form. There are no strong dependencies between the different profiles, other than the need for data descriptions to have access to controlled vocabularies where required.

The CDIF metadata will indicate which Core Profiles are supported using dcterms:conformsTo statements, as described in the 'Implementing CDIF Profiles' section, below. The JSON-LD expression of the metadata can be published as a stand-alone file or embedded in a landing page, as appropriate.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The Discovery profile covers static data files and queryable services. The Access profile uses ODRL[44] in combination with user-defined Actions, as described below. This requires also a minimal use of DCAT[45], to identify the object for which access is being described. Controlled vocabularies - the semantic component - are described using SKOS[46], and these can be published for use independent of a data set, or alongside the description of data. (If an OWL[47] expression of an ontology exists, that can be referenced from the SKOS.) Data description assumes that the data is available in a text-based form, for the purposes of the current draft. This includes common formats such as CSV[48], other delimited formats, and fixed-width formats. (Support for other types of data is anticipated in future.) The description of data relies on a subset of the DDI-CDI specification[49].

This core set of profiles addresses the highest priority issues that have emerged through the WorldFAIR project and the consultations of the CDIF Working Group. Of course, to fully support cross-domain FAIR many important areas remain to be covered. The diagram below helps describe the current plans for future CDIF profiles.

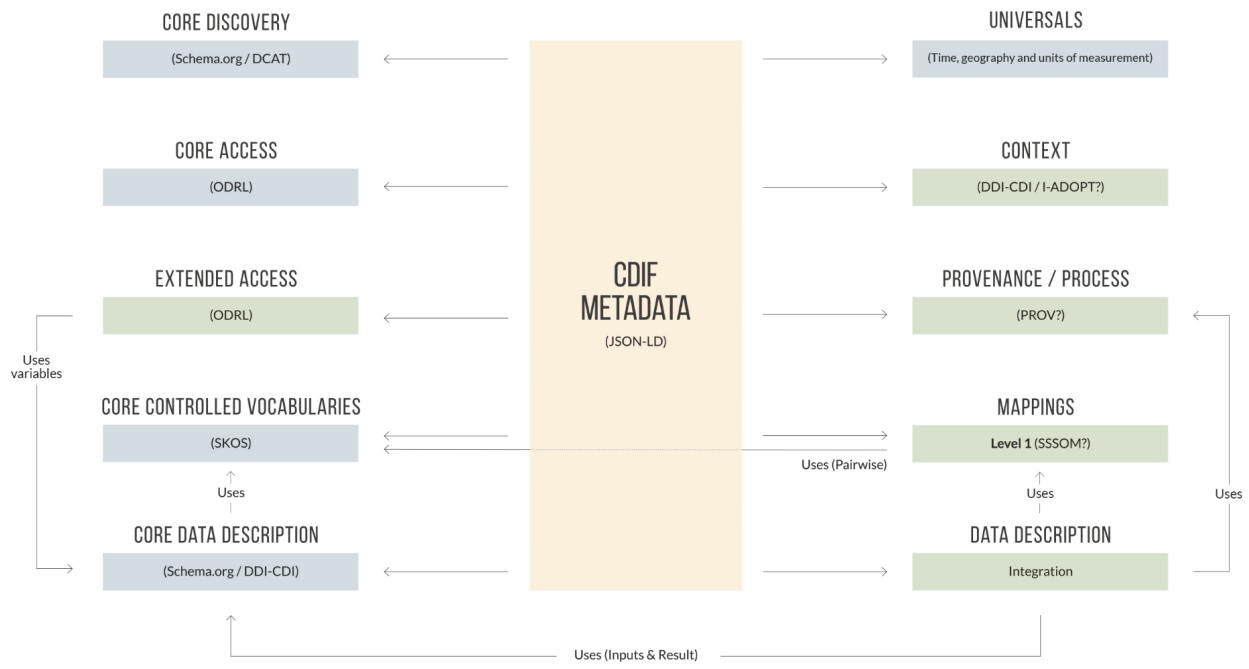

Figure 1b. Potential Future CDIF Profiles.

[44] https://www.w3.org/TR/odrl-model/
[45] https://www.w3.org/TR/vocab-dcat-3/
[46] https://www.w3.org/TR/skos-reference/
[47] https://www.w3.org/TR/owl2-overview/
[48] Comma-separated values
[49] https://ddialliance.org/Specification/ddi-cdi

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Profiles anticipated in future appear in green, and are discussed in Section 7. These include a larger number of dependencies between profiles as a consequence of the information they contain. Describing an integration of data sets requires the description of each of the input data sets and the output data, as well as information regarding the mappings employed and the process by which the integration was performed. Mappings require not only the correspondences, but also a description of the resources being mapped - in this case, controlled vocabularies. Mappings themselves are a focus topic within the RDA FAIR Mappings group, with which CDIF intends to align, so a "Level 1" label has been applied. The Simple Standard for Sharing Ontology Mappings[50] (SSSOM) is widely viewed as useful, but it is also expected that additional specifications will be required.

As CDIF develops further, this picture might change, but at least indicates some of the new profiles which are likely to be included.

6.1. Design principles

6.1.1. Overview

The Cross Domain Interoperability Framework is based on a set of concrete, implementation-level principles intended to enhance any community's data management approaches and reduce the barriers to reusing data across domains. These principles will guide the selection of CDIF recommended standards and approaches intended to inform project-level planning and to provide practitioners with clarity regarding the work involved and the skills, personnel, and technology required. The guidance offered by these principles should not be read as absolute requirements; they define aspirations that should be pursued as far as feasible. In the discussion that follows we will use the term 'data' for any structured information representation that is intended for use by computer information systems. This includes both direct observational or model data and data describing other data, commonly differentiated as metadata, but subject to the same principles.

6.1.2. List of principles

In this section, we set out the principles that have guided the development of the CDIF to date.

6.1.2.1. Pragmatism

CDIF is not just a set of theoretical standards and guidelines, but a recommended approach that can realistically be implemented, taking into account the capabilities and limitations of (meta)data producers and consumers. This implies considering the known time limitations, resources, and technical capabilities of those expected to implement it. The framework must balance between flexibility and rigid requirements, providing helpful guidance and a systematic approach, with room for adaptation to new situations and requirements, but within clearly defined bounds to preserve interoperability.

[50] https://mapping-commons.github.io/sssom/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.1.2.2. Mainstream

Technology and standards recommended by CDIF should build on existing systems and the legacy investment of (meta)data providers and consumers. User familiarity with recommended implementations is desirable, and these should be widely recognised and in line with current production systems. There will always be new standards and technology, and CDIF will need to keep pace with these on an implementation level, but the goal should always be to employ existing technology, rather than to try to anticipate what practice will be in the future. Use existing Web architecture[51] whenever possible. For example:

- HTTP methods (POST, GET, PUT, PATCH, and DELETE) to operate on resources;
- Resource-oriented, API-driven architecture to support file-based download as well as interactive, query-driven access;
- Ensure identifiers are dereferenceable over the Web using the HTTP(s) protocol;
- Alternate serialisations can be requested using content negotiation or signposting links.

6.1.2.3. Atomicity

Manage data at the finest granularity possible. Information is organised into entities that have sets of properties, and each property has a value that is an instance of another entity, or a simple value that is a number or a category (represented by a word). In practice, this can lead to complex, nested data schemas that are difficult to understand, manage and use, so it is necessary to consider the requirements of the application for which the information will be used, and the trade-off between the cost of complexity and the benefits of granularity. Information captured in free text is useful for human reading, and natural language processing is getting better at extracting structured data from text, but it is subject to semantic fuzziness and misinterpretation. The benefit of highly granular, or 'atomic' data is composability: it can be disaggregated and reaggregated to accommodate diverse and evolving standards, and support projection into different forms in an integration-on-demand model.

6.1.2.4. Data as an asset that is worth investment

View data as an asset that has value outside of its immediate application. In practice, this is implemented by making data and associated documentation (metadata) persistently available on the Web, with identifiers at all levels of granularity that can be dereferenced using standard Web architecture, and representation formats using one or more prevailing domain-specific or domain-neutral standards. Identifiers in the data should provide linkage to other resources on the Web.

The goal is to enable automation of the data-discovery-to-use workflow to the maximum extent possible. CDIF is intended for data producers who put data on the Web with the intention that it be used. The approach is to generate machine-actionable (atomic) documentation that is as expressive as possible, e.g., very precise and clear definitions, complete descriptions, and links to supplementary metadata. CDIF can be

[51] https://www.w3.org/TR/webarch/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

understood as recommendations for implementation of the FAIR Data Principles, extending them to address more concrete issues.

Documentation for data should be targeted to future users, human or automated, who are temporally or socially removed from your current community. Background knowledge and assumptions that are currently obvious might not be obvious to those users. The terminology used should be clearly defined in plain language, and the definitions should be updated as needed such that they will be accessible and current for the lifetime of the data.

6.1.2.5. Flexibility and tooling to encourage uptake

Create and share the ability to project data and its metadata into serialisation formats that are useful for a variety of data consumers across domains. Offering data content in multiple human languages supports different language-based communities. From a machine-processing point of view, multimodality will usually entail:

- Creation, documentation, exposure, and maintenance of authoritative mappings between syntactic and semantic conventions within and between domains;
- Development of software that leverages these mappings to automate the presentation of data in forms familiar to end users or to developers.

Standards used to interchange data should be as domain-agnostic as possible, with mapping files and documentation, which must qualify any loss of precision where it occurs. In practice, this means that categorical properties should usually be represented using key-value approaches, with conventions for vocabularies specifying both property types, and vocabularies specifying the range of values associated with each property.

6.1.2.6. Broadcast system capabilities

The Web is too big for all users to keep up with who does what, who holds what data, and how to access the data. Data providers have to be proactive by presenting a machine-actionable description of their interoperable data offerings and services, defining the datasets offered, service protocols used, vocabularies, access policies, or any other digital resource. The CDIF needs to define conventions for publishing server capabilities.

6.1.2.7. Robustness

Trust in data quality and the longevity and reliability of systems in a cross-domain digital ecosystem is critical to user engagement. Maintaining this robustness has a technology aspect and a business aspect. The architecture and implementation of systems in a cross-domain digital ecosystem should support the addition, update, or retirement of their components. This entails avoiding dependence on any single component, whether it be hardware, software, standards, or conventions. A component-abstraction layer in the architecture can enable of-the-moment components to be swapped. Focus on maintaining basic

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

cross-domain interoperability functions using proven off-the-shelf components. Functionality can be extended, but beware of becoming a silo through dependence on bespoke tooling or approaches.

Robustness from a business point of view is in many ways a bigger challenge. We can point at some necessary factors, but these are not always sufficient. Systems must define their market niche, making sure that there is a value proposition that is clear to the system stakeholders so they 'buy in' and continually support the system. Effort to increase awareness of the system capabilities — marketing, training, workshops — is likely to be necessary. Avoid building cross-domain technology 'just because': make sure someone clearly benefits from the technology, and they know about the technology.

6.1.2.8. Show and tell

Collect and publish metadata related to the use of CDIF to show what implementers have done and what the results of those investments have been. A living, visible system will generate trust and uptake; and ensure that the CDIF remains fit for purpose as technology and priorities evolve. A framework that is not used will not continue to function. Make it easy to determine what standards are used, where data are shared, how data can be accessed, who is reusing data, and what the effects are of data reuse. This kind of information will also facilitate changes to improve the framework and its implementation.

6.2. Technical expression of metadata

The Internet and World Wide Web make an immense body of data, information, and knowledge accessible to anyone with a computer that can interact with the system. The Web can be thought of as a library containing a large fraction of the written, recorded, and graphic works of humanity, or as a database containing an almost unimaginable spectrum of data from scientific research, historical records, government records, sensor networks, etc.

The challenge with this mountain of resources is to find particular nuggets useful to answer questions and solve problems. Addressing this problem requires, firstly, understanding what kinds of things the World Wide Web offers, referred to as 'resources' in this discussion. Secondly, how are these resources described and documented such that they are useful? Thirdly, how can descriptions of resources be made available on the Web to guide users to find and use the resources?

6.2.1. Digital Objects

A resource is an identifiable thing of interest to someone; it might be a Digital Object (DO) or a Non-digital Resource. A Digital Object is a packaged, identifiable sequence of digital bits that carries some information. A Digital Object has exactly one digital representation: a bitstream. The Digital Object bitstream might be the resource of interest, or it might be a representation of an abstract or physical resource that cannot be transmitted electronically (see HTTP Range-14[52]). The identifier for a Digital Object can be dereferenced to access the object directly. A non-digital resource is a material entity (e.g. person, rock sample), an abstract

[52] https://en.wikipedia.org/wiki/HTTPRange-14

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

entity (e.g. Donald Duck, The Land of Oz), or a 'Work' or 'Expression' in the FRBR sense[53] (e.g. Beethoven's 9th Symphony, Dickens' 'Tale of Two Cities'). Identifiers for non-digital resources must dereference on the Web to a Digital Object that is a representation of the non-digital thing and can be transmitted electronically.

In the FAIR Digital Object Framework (FDOF), a Digital Object[54] (DO) is a specific bit stream that carries some information and has a persistent, registered, resolvable identifier (PID) that can be resolved to obtain a PID kernel record providing documentation for the source of the PID, expected lifetime, type of resource it identifies, linkage to the resource it identifies, and other attributes specified in a schema identified in the PID kernel record (PID profile)[55]. The PID kernel record is a metadata record conforming to a particular PID profile.

Digital Objects are FAIR (FDOs) when they are part of an ecosystem comprising services and infrastructure to support realisation of the FAIR (Findable, Accessible, Interoperable, Reusable) principles. In the FAIR Digital Object Framework (FDOF) there must be a mechanism to access either the object or its metadata by dereferencing the object's PID. Metadata content must enable the identified resource to be found, used and cited, enable interoperability and reuse, and include machine-actionable statements about dependencies and licensing[56]. Bonino et al. (2022)[57] propose some approaches to access the FDOFIdentifierRecord (Kernel metadata) and other FDOF requirements; the level of adoption for this approach is uncertain.

6.2.2. Metadata

In order for resources to be discoverable on the Web, the search applications that are used to find things must locate some representation of the resource and must be able to parse that representation and generate indexes for searching. In the realm of linked HTML Web pages, search engines parse the text content and links on Web pages to create text-based indexes and use links to find other pages to crawl the Web. This approach does not work for datasets, images, sound recordings, videos, and other non-narrative text resources, so separate representations of their content are constructed as metadata, in a format that can be parsed and indexed by search applications. This pattern is also applied to text resources to provide more explicit documentation. At the simplest level the content of the metadata can consist of text describing the resource and a link to access the resource, analogous to what is included on the cards in a legacy library card catalogue. Representing the metadata content using a structured, machine-readable format makes the information more precise and accessible to software agents.

6.2.3. Metadata content

In the digital world, a wide variety of metadata schemes have evolved for describing resources. These schemes are structured to allow a richer understanding of the information, and typically at least include information about the set of fifteen generic elements identified as the Dublin Core: Creator, Contributor,

[53] https://www.loc.gov/cds/downloads/FRBR.PDF
[54] Note that 'Digital Object' is capitalised in this document to emphasise that it is being used in this specific sense.
[55] Weigel, T., Plale, B., Parsons, M., Zhou, G., Luo, Y., Schwardmann, U., Quick, R., Hellström, M., Kurakawa, K. (2018). RDA Recommendation on PID Kernel Information (Version 1). DOI: https://doi.org/10.15497/RDA00031
[56] https://fairdigitalobjectframework.org/
[57] Bonino, L., Guizzardi, G., Sales, T., (2022) 'FAIR Digital Object Framework Documentation', https://fairdigitalobjectframework.org/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Publisher, Title, Date, Language, Format, Subject, Description, Identifier, Relation, Source, Type, Coverage, and Rights, first drafted at a 1995 meeting in Dublin, Ohio[58]. These elements are defined at an abstract level and served well with free text content values for use by humans. Such semi-structured metadata is insufficient to support machine-actionable reuse of the described resources.

In order to be machine-actionable, the structure, syntax, and element-value representations in a metadata document must conform to conventions that client software can be programmed to parse and 'understand'. 'Understand' in this context means recognise the incoming bitstream content and take appropriate, useful action. The metadata provider must communicate the conventions used to serialise the metadata they provide. Ideally this is done with an identifier for a specification document that details the conventions used. Some widely used metadata specifications include DCAT, DataCite Schema[59], ISO 19115-1[60], EML[61], FGDC CSDGM[62], CERIF[63], schema.org, and DDI. These specifications determine the structure and syntax of metadata documents, but leave latitude on how the values of some metadata elements are represented, and often offer multiple valid approaches to representing the same information.

6.2.3.1. Metadata profiles

Achieving the level of metadata interoperability required for CDIF will require the adoption of one (or a small number of) metadata specification(s), along with more specific conventions on vocabularies used for metadata properties. We refer to such a set of specific conventions as a profile. CDIF provides recommendations for a metadata profile compatible with machine processing. This profile includes the base file MIME-type, an information model for the metadata content, and how that information is represented both syntactically and semantically. Most profiles are based on an existing metadata specification, e.g. schema.org, DCAT, ISO19115-1, EML, DDI-CDI, but provide additional detail to resolve ambiguities in the base specification, or rules for vocabularies and data types for element values that extend or restrict the base specification. The simplest presentation of a profile specification can be a text document that describes the information required, identifies the base specification, and states any conventions or rules for profile conformance. Such a document could be used by a software developer writing code to use information in metadata conforming to the profile. Profiles might also be specified in a machine-actionable way, e.g., the Dublin Core Tabular Application Profiles (DCTAP)[64], Profiles Vocabulary[65], SHACL rules[66], XML schematron rules[67], or other schema or rule representations. Using a rule-based representation for metadata profiles provides an approach to defining and communicating metadata constraints that can be validated automatically to support metadata profile interoperability, reusability, and quality.

[58] https://www.dublincore.org/resources/metadata-basics/
[59] https://schema.datacite.org/
[60] ISO 19115-1:2014 Geographic information — Metadata https://www.iso.org/standard/53798.html
[61] https://eml.ecoinformatics.org/
[62] Content Standard for Digital Geospatial Metadata (CSDGM) https://www.fgdc.gov/metadata/csdgm-standard
[63] https://eurocris.org/eurocris_archive/cerifsupport.org/cerif-in-brief/index.html
[64] https://www.dublincore.org/specifications/dctap/
[65] https://www.w3.org/TR/dx-prof/
[66] https://www.w3.org/TR/shacl/
[67] https://www.schematron.com/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.2.4. Making metadata accessible

The use of standardised metadata to describe resources is one facet of discoverability, but mechanisms to make these metadata descriptions discoverable by search applications or other machine agents is also necessary. Search applications harvest metadata to index and present in search results. In other cases a machine agent might have an identifier for a resource on the Web and want to get its metadata to learn about the content and capabilities of the resource.

6.2.4.1. Signposting

Signposting is an approach to discovering the content and capabilities of resources accessed by resolving URIs on the Web. Like the FAIR Digital Object framework, it starts with an identifier that can be resolved, and uses typed Web links (IETF RFC8288[68]) and IANA registered relationship[69] types to enable an agent to discover what is identified and navigate to metadata with more information. These signposting links can implement the linking requirements of the Fair Digital Object Framework[70].

The CDIF metadata requirements outlined below include recommended or optional properties necessary to create a FDO PID Kernel information record (FDOFIdentificationRecord in Figure 1, above) as recommended by the Research Data Alliance (RDA)[71]. The mapping from the PID Kernel information record to the CDIF metadata schema.org recommended implementation is shown in Appendix A. 2. Mapping from Signposting link relation types to CDIF metadata elements is shown in Appendix A. 3. Thus, Signposting can be used to implement FDOF requirements, and in the following implementation discussion we outline how Signposting is compatible with CDIF.

[68] https://datatracker.ietf.org/doc/html/rfc8288
[69] https://www.iana.org/assignments/link-relations/link-relations.xhtml
[70] https://www.slideshare.net/hvdsomp/fair-signposting-a-kiss-approach-to-a-burning-issue
[71] Weigel et al. (2018). https://doi.org/10.15497/RDA00031

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

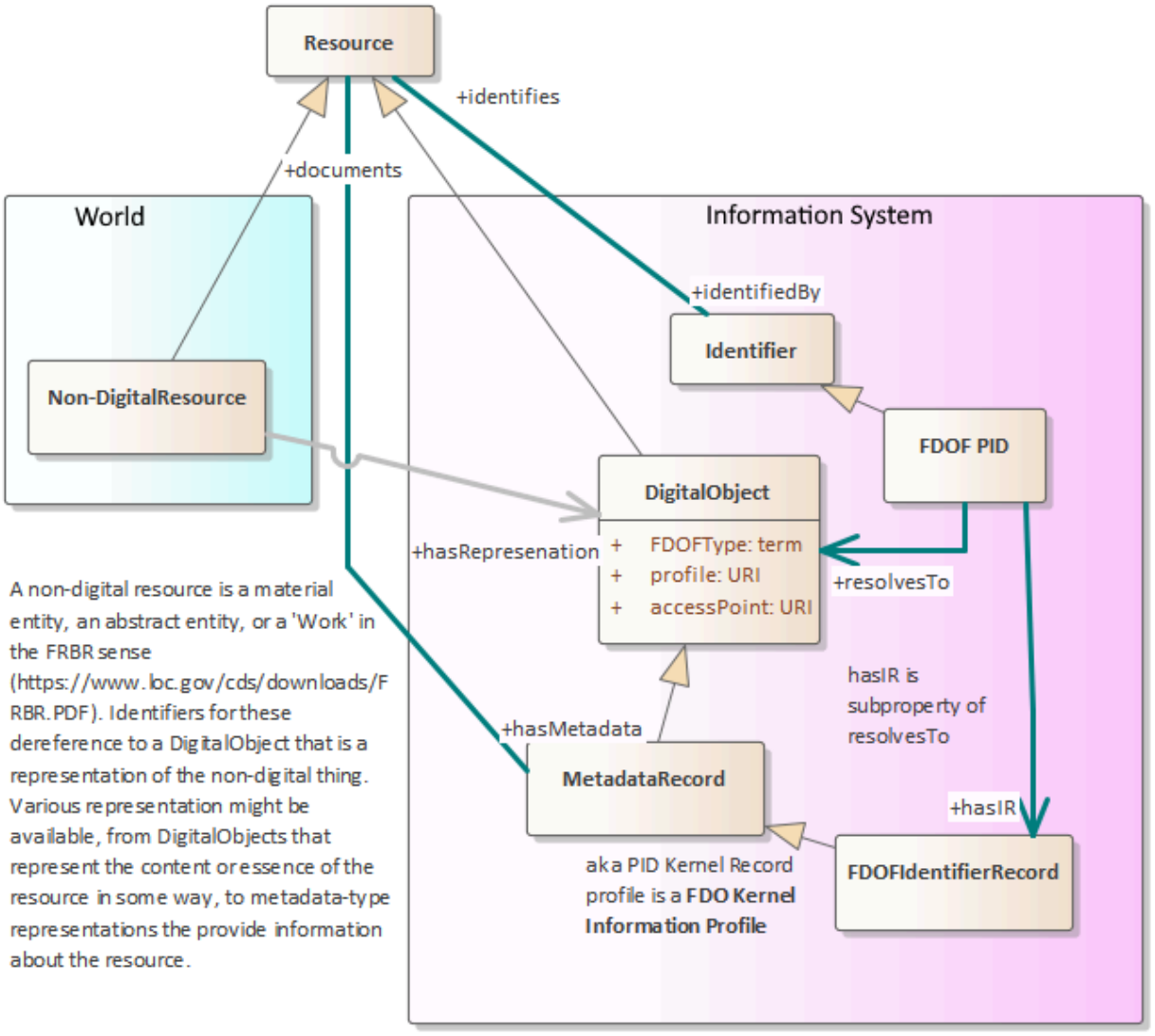

Figure 2. FDOF-CDIF metadata relations.

6.2.4.2. Search engines

One of the drivers for the success of the World Wide Web is the emergence of search applications that use various mechanisms to traverse the Web, find Digital Objects of interest, analyse those objects to extract information about their content, and create indexes that support user searches for resources of interest. Creators and providers of Web resources (publishers) seek to increase visibility and usage of their published resources by assuring that search engines (aggregators) find and index their products. This section outlines various approaches used to make metadata accessible to search engines.

Finding documents to index: Web-crawling is still an important approach to finding and indexing resources on the Web, and this approach is supplemented by Signposting. A different and widely-used approach is the sitemap, which is a list of Web locations (URLs) for files that a hosting agent wants search engines to index.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Many search engines enable providers to register sitemap locations. Alternatively, a widely-used convention on Web servers is to place a 'robots.txt' file in the root directory of a Web site. This file contains links that point to one or more sitemaps that should be indexed.

Getting the metadata: Once a crawler for a search application finds a document that should be indexed, it must determine if there is structured metadata to index, and what conventions the metadata uses. Possible approaches:

1. Each resource has an HTML landing page that describes the resource for human users, and contains embedded metadata for machine clients. Metadata can be embedded in landing pages using the HTML <script> element, in alignment with the Data on the Web Best Practices, specifically section 8.2, Metadata[72] (See Example 1). This approach requires that each published resource has a human-readable landing page, intended to be the target of search by human users. Scripts are normally embedded in the <head> section of an HTML document. The <script> element has at minimum a 'type' attribute that provides a MIME-type specifying the type of script.

```
<html xmlns="https://www.w3.org/1999/xhtml">
  <head>
      <script type="application/ld+json">     {
            "@context": {"@vocab":"https://schema.org/"},
            "@type":"Dataset",
            "@id":"https://doi.org/10.17616/R31898",
            ....    }
      </script>
    ....
  </head>
  <body>
.....
  </body>
</html>
```

Example 1. A JSON-LD metadata object embedded as a script in an HTML document.

2. Metadata can be embedded in the HTML <head> section of a landing page using HTML <meta> elements, which have a 'name' attribute that can be used to identify different metadata properties (see Example 2). This approach has been implemented by some off-the-shelf repository software (e.g. Dataverse). The HTML <meta> elements are intended to describe the HTML document that contains the <meta> element[73], not some external resource that the Web page is about. CDIF recommends against this approach and suggests using the script approach (No 1 in this list) instead because that is more widely used and allows richer metadata content to be included.

[72] https://www.w3.org/TR/dwbp/#metadata

[73] https://www.w3.org/TR/2011/WD-html5-author-20110809/the-meta-element.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
<html lang="en">
    <head>
        <meta name="DC.creator" content="Richard, S.M." />
        <meta name="DCTERMS.dateAccepted" content="2018-11-10T02:57:04Z" />
        <meta name="DCTERMS.available" content="2018-11-10T02:57:04Z" />
        <meta name="DCTERMS.issued" content="1999" />
        <meta name="DC.identifier" content="http://hdl.handle.net/10150/630821" />
        <meta name="DC.description"  content="This map documents t..." />
        <meta name="DC.language" content="en" />
        <meta name="DC.publisher" content="Arizona Geological Survey (Tucson, AZ)" />
        <meta name="DC.subject" content="Arizona Geological Survey Open File Reports" />
        <meta name="DC.subject" content="La Paz County" />
        <meta name="DC.subject" content="Harquahala Mountains" />
        <meta name="DC.subject" content="White Marble Mine" />
....
    </head>

    <body>
....
    </body>
</html>
```

Example 2. HTML meta tags with metadata about a resource.

3. Metadata can be linked from the landing page using the HTML <link> element in the <head> section to provide a Web locator (URL) that can be used to retrieve a full metadata document about the described resource. The link element has the "rel='describedby'" attribute to indicate that the link is to metadata; a 'type' attribute to provide the MIME type of the target metadata record; and optionally a 'profile' attribute to identify specific metadata conventions. Note that multiple links could be provided to different metadata views, e.g. a CDIF record, a PID Kernel record, a FDOFidentifierRecord, etc. This approach depends on Web-crawlers to identify and follow these links to get the metadata that the provider wants indexed. This is one of the Signposting approaches[74].

4. The server providing the resource can be configured to include <link> elements in the HTTP response header[75] that indicate the location of machine-actionable metadata describing the subject of the URL target. As in the HTML link approach, these links would have a "rel='describedby'" attribute to indicate that the link is to metadata, a 'type' attribute to provide the MIME type of the target metadata record, and optionally a 'profile' attribute to identify specific metadata conventions. The advantage of this approach is that the HTTP response header links can be provided for any resource that has an HTTP URL, so links to metadata can be accessed for non-textual resources that do not have an associated HTML landing page. If the download size for the resource is large, a client can use

[74] https://signposting.org/
[75] https://tools.ietf.org/html/rfc8288

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

the HTTP 'head' request to access this header information without downloading the entire file[76]. This approach would enable indexing of large resource collections that have a single landing page, but for which the individual resources do not have a landing page, e.g. a STAC catalogue. The downside is that many client applications do not use the HTTP header information[77]. This is a second Signposting approach. CDIF recommends this approach for resources that do not have landing pages; see implementation recommendations below.

5. A sitemap can point directly to metadata documents in formats that the search engine can parse. With the basic sitemap XML schema, all metadata would need to conform to a single profile. In the implementation section below, the CDIF proposes using an extension to the sitemap scheme that allows labelled links to the indexing targets.

6. Another option is for the sitemap to provide a URL that retrieves a document containing a collection of metadata records, something like the US Government Data.gov Project Open Data Catalog[78], or Ocean Info Hub graph first approach[79], with individual records providing CDIF profile metadata.

How do harvesters know where to look?

- Publishers register metadata services with a harvester, e.g. by providing a URL to GET a sitemap or other metadata catalogue document, e.g. an OpenGeospatial Consortium (OGC) Record collection (catalogue).[80]
- Server robots.txt has link to sitemap.xml file; The sitemap.xml lists Web locations that a crawler should index.

Once the harvester has a URL for a location to index, how do they know where the metadata is relative to that location? Possible approaches:

- Try a HTTP HEAD request[81] on the URL and inspect. If the Content-Type header value is a known metadata type and profile, then the URL will get a document containing a single metadata record that can be indexed. Failing that, the client can look for links in the HTTP response header; if there is a link with rel='describedby', with a known type and profile, get the content at that link.
- GET the content at the URL. Look for a <script> element with a known type and profile. Failing that, look for elements with rel='describedby' and a known type and profile, then get the content at that link.

This general procedure can be simplified if the sitemap or other catalogue the harvester is iterating through provides labelled links.

What does the harvester do with the metadata?

[76] https://tools.ietf.org/html/rfc7231#section-4.3.2
[77] https://www.w3.org/TR/ldp-bp/#use-case-2-providing-metadata-in-both-http-headers-and-html-body
[78] https://project-open-data.cio.gov/v1.1/schema/catalog.json
[79] https://book.oceaninfohub.org/indexing/graphpub.html
[80] https://docs.ogc.org/DRAFTS/20-004.html#sc_record-collection-overview
[81] https://www.rfc-editor.org/rfc/rfc9110#HEAD

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

There are many possible approaches a client application could use to extract the information it needs from a metadata record. The simplest and likely most accurate approach is for the metadata to conform to a profile that the application is programmed to parse, and to communicate that profile conformance to the application. This entails two requirements. The profile must be documented in a way that allows software developers to write code to parse metadata conforming to the profile, and the profile must have an identifier that can be used to assert conformance.

- The use of <script> or <link> elements (in the HTTP or HTML header) allows metadata to be offered following multiple specifications with the 'type' and 'profile' attributes used to identify the particular conventions[82] (Example 3).

```
<head>
    <script type="application/ld+json;type=BlueSky1.17">
          "@context": {"@vocab":"https://schema.org/"},
          "@type":"Dataset",
          ....    }
    </script>
  ....
```

Example 3. Script with a type parameter in the MIME-type string

Minimally, the metadata record should assert the specification used to generate the record in a metadata property.

6.3. Discovery

The core of the CDIF profile for resource discovery is a set of implementation-independent content requirements that specify the required information to support a basic level of discovery interoperability for resources of any type. A recommended metadata implementation for these requirements using the schema.org vocabulary and JSON-LD is outlined in Appendix 1. Next, approaches to make the metadata findable are recommended for resource publishers. A following section outlines recommendations specific to resources distributed via interfaces that allow users to select specific parts of a resource or to request some processing to obtain a customised subset or representation.

6.3.1. Metadata content requirements

The following list includes the minimum required content for basic resource description, discovery, and access. This recommendation is a synthesis of various metadata schemes, including ISO 19115-1:2014, schema.org conventions from ESIPFed Science on Schema.org and Ocean Data net, DCAT, DCAT-AP, and FDO Kernel Attributes-2.0[83]. A mapping between these various schemas and CDIF content elements is available in

[82] https://www.w3.org/TR/dx-prof-conneg/#dfn-profile

[83] https://docs.google.com/document/d/1OF49wTNVuv-6OXlNerhBTqVtHyc7jutTaUHjn6BZCs0

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

a Google Spreadsheet[84]. Note that these content requirements are scoped for a broad spectrum of resource types. It is expected that other fields will need to be added in extensions for specific kinds of resources.

6.3.1.1. Required

If the content of a required element does not provide useful information, the metadata is considered useless for even the most rudimentary discovery use cases. Conformant metadata MUST provide valid values, i.e., a meaningful title that identifies the resource, either a URL or text statement of how to obtain the resource, a statement of any licensing, usage, or access constraints (i.e., Rights), and identifiers for the specification of the metadata serialisation and the type of the resource described.

- **Resource identifier (1 entry)**: A globally unique, resolvable identifier for the resource described by the metadata record.
- **Title** (**1 entry**): Succinct (preferably <250 characters) name of the resource; should be sufficient to uniquely identify the resource for a human user.
- **Distribution: URL, Distribution object,** or **Access Instructions** (**1 entry**): If the resource is a digital object accessible online, provide a URL that will retrieve the resource. If the resource has multiple representations, provide a Distribution Object documenting the various options with a URL and representation profile for each. Metadata for distributions through an API that allows query, filter, or processing as part of a data access request are described in the Queryable Distribution Interfaces (API) section, below. If the resource is not accessible online, provide a URL to a landing page used to access the resource, or minimally, provide a text description explaining how to access the resource in the metadata (Access Instructions).
- **Rights** (**1 to many entry**): Information about required access permissions, licences, contractual requirements, use constraints, and security constraints. Might be described in text or through links to external documents. (See 6.4. Data Access for providing machine-actionable rights descriptions.)
- **Metadata profile identifier** (**1 to many**): Identifier for metadata specification (profile) used to create this metadata record. Generally this will be populated automatically if the metadata is created using CDIF aware tools.
- **Resource type** (**1 to many**): A scoped name (label with classification scheme) that specifies the kind of resource described by the metadata. The resource type might be used to determine validation requirements specific to descriptions for that kind of resource.

[84] https://docs.google.com/spreadsheets/d/1wFuJ4RRlNirnrPfuY_d57I9_pnaNibw4nltNTkruSp0, adapted from https://doi.org/10.15497/RDA00069, which is based on An Analysis of Crosswalks from Research Data Schemas to Schema.org: https://doi.org/10.1162/dint_a_00186

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.3.1.2. Required, but nilable

These are content elements for which every resource should have useful information, but for which the information may not be available. A corresponding field should be included in each metadata record, but may have value 'nil:missing', 'nil:unknown' or similar nil value. Use 'nil:notapplicable' for Temporal Coverage, Geographic Extent or Statistical Variable when these are not applicable to the described resource.

- **Description** (1 entry): Inform the reader about the resource's content, context, provenance, and any other information deemed useful for future cross-domain usage.
- **Originators** (1 to many entries): One or more parties (person or organisation) that have a role related to the origin of the resource, e.g., author or editor. Each party has a name (label), identifier, and optional contact information.
- **Modified Date** (1 entry): Date (not temporal extent) when the most recent changes to the resource were completed. Use a "year" or [ISO 8601 date and time] format. Alternative date formatting must be machine-readable and consistent across all datasets.
- **Distribution Agent** (1 entry):The party (person or organisation) to contact about accessing the resource. Each party has a name (label), identifier, and optional contact information. If there are multiple distribution options with different contact points, the Distribution Agent should be specified as part of the Distribution Object.
- **Statistical Variable** (1 to many entries): Only applicable to datasets, otherwise nil:{reason}. A complete description of a dataset should include a list of the fields in the data, with each field mapped to a variable that is represented by the content of that field. (See 6.8 Implementing CDIF profiles for more information.) Variable definitions should minimally specify the property represented by name. Identification of the property represented with a resolvable URI is strongly recommended. Variable descriptions should include documentation for how units of measure and reference systems for values are specified (see 6.7. Universals: Time, Geography, and Units of Measurement). Details of data structure and schema more closely related to interoperability, data integration, and usage than to data discovery are discussed in 6.6.5. Describing Data to Make it “Integration-Ready”.
- **Temporal Coverage** (1 entry). The time interval represented by or the subject of the described resource. This could be the time interval when data were collected, or an archaeological or geological time interval that is the subject of the resource. Need to account for clock time, calendar time (Gregorian, Julian, Hebrew, Islamic, Chinese, Mayan...), cyclical time (summer, first quarter, mating season, new moon, pay day) and for named time ordinal eras (Jurassic, Younger Dryas, Early Minoan I, Late Stone Age). See OWL Time[85].
- **Geographic Extent** - horizontal (if applicable, 1 entry, minimum bounding rectangle or point): In order to support cross-domain searches based on geospatial location, location coordinates must be given in decimal degrees using the WGS 84[86] datum. There are various other systems for describing location (see 6.7.2. Space); these can be provided as alternate location descriptions, recognizing that they might be

[85] https://www.w3.org/TR/owl-time/
[86] https://en.wikipedia.org/wiki/World_Geodetic_System

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

meaningful to some metadata harvesting agents. Some resources may not be usefully described by a WGS 84 extent, in which case indicate nil:notapplicable; this would include extraterrestrial resources.

- Bounding Rectangle: **North Bounding Latitude, South Bounding Latitude, East Bounding Longitude, West Bounding Longitude**. The minimum rectangle that completely contains the coverage extent for the resource content. Coordinate order and syntax are determined by the serialisation profile.
- Point: **Latitude, Longitude**. A centroid point for the coverage extent of the resource, or the location of the resource content if a point location is appropriate. Coordinate order and syntax are determined by the serialisation profile.
- Named location: Place name referenced to some gazetteer. Use scoped name pattern {label, authority, optional identifier} (see [(placename, Gazetteer (, date))].

6.3.1.3. Required for metadata management

These elements provide essential information for the operation of a distributed catalogue system with harvesting of metadata between catalogue servers. Values should be populated automatically by metadata creation tools, requiring no user input. Nil values are allowed.

- **Metadata Date** (1 entry): Last metadata update/creation date-time stamp in ISO 8601 date and time format. This may be automatically updated on metadata import if a metadata format conversion is necessary.
- **Metadata Contact Agent** (1 entry): The party responsible for metadata content and accuracy; Agent object includes a name (label), identifier, and optional contact information
- **Metadata Identifier** (1 entry): The identifier for the Digital object that contains the metadata.

6.3.1.4. Recommended

Other properties that should be specified if possible and relevant. All are optional.

- **Checksum**. (0 or 1): A string value calculated from the content of a digital object that allows verification that the content of the object has not been modified. Even insignificant changes to the content of the file will change its checksum. The algorithm used to calculate the checksum must be documented. See also RFC-6920 'Naming things with hashes'[87] that establishes ways to identify checksum algorithms and to represent checksum values as a URI. Note that checksums apply to specific digital objects, typically a unique resource representation. Non-digital resources do not have checksums; their representations can have checksums. See implementation notes in Appendix 1.
- **Funding**. (0 to many entries): Cite funding sources (Grants, contracts...). Each source has a grant or contract identifier, source organisation, and label.

[87] https://www.rfc-editor.org/rfc/rfc6920.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- **Keyword** (0 to many entries): Distinguish 'tags' and 'controlled terms'. Tags are simply words that a metadata creator thinks will be useful for users to identify resources of interest. Controlled terms are words defined in a vocabulary that minimally include the word (a fixed string to identify the term for humans) and a definition. Each term represents some concept. More semantically rich vocabularies would include resolvable identifiers, source information, and links to related terms (see Cox et al., 2021[88]). One common set of relationships in a vocabulary is a kind-of hierarchy linking broader to narrower concepts. Controlled terms should minimally be represented in metadata with a label and scheme name that identifies the source vocabulary; ideally a term URI and scheme URI could be included for more accurate identification and data integration.

- **Policies** (0 to many entries): Policies used in management of the described resource, including whether the content may be changed (mutable or immutable), any scheduled updates, what is the expected lifetime for resource availability, what (if any) is the maintenance schedule, versioning, documentation for changes and change requests. Explicit support for specific policy frameworks can be included (e.g., CARE).

- **Publication Date** (0 or 1): Date (not temporal extent) when the resource was made accessible. Use a 'year' or ISO 8601 date and time format. Alternative date formatting must be machine-readable and consistent across all datasets. If no publication date is known, estimate the publication date range, enter the oldest year as the publication date, and include the estimated date range in the Description field.

- **Other related agents** (0 to many entries): Recognition for others who have contributed to the production of the resource but are not recognized as authors/creators. Includes a variety of roles like maintainer, publisher, point of contact, copyright holder, contributor (see e.g. DataCite contributor types[89], ISO19115-1 role code[90])

- **Related resources** (0 to many entries): Links to related data, publications, annotation, data sources, software used, etc. Links have at least a label, relationship type, and resolvable target resource identifier.

- **Version** (0 or 1): If the resource is versioned, specify the label for this version. Version labels should follow a scheme that allows alphanumeric sorting reflecting the order of version release.

6.3.2. Queryable distribution interfaces (API)

Many resources are accessed using interfaces that allow customization of the resource representation and content that are delivered to a user. These interfaces are referred to as an API or 'Application Programming Interface', which is a specification of how to interact with a machine agent. Such a specification typically requires a communication protocol, description of the functions offered, and description of the content of messages transmitted between a client and the agent offering the interface. The focus of CDIF discovery is

[88] https://doi.org/10.1371/journal.pcbi.1009041
[89] https://datacite-metadata-schema.readthedocs.io/en/4.5_draft/properties/recommended_optional/property_contributor.html#a-contributortype
[90] https://wiki.esipfed.org/ISO_19115_and_19115-2_CodeList_Dictionaries#CI_RoleCode

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

on communication protocols that use the internet and Hypertext Transfer Protocol (HTTP), which is the technology that underlies the World Wide Web. In this case the communication protocol can simply be specified as HTTP (or HTTPS). Various other protocols exist and new ones will be invented, these are out of scope for this recommendation.

Messages from a client to a service provider include, for example, requests to invoke some operation, requests to get content, and messages requesting information or providing information as part of a workflow (e.g. authentication, job status, inputs to configure processes). Messages from the service provider to the client include, for example, information about the service capabilities, status responses to requests (success, error..., in progress), or requested content or process results. The content of these messages uses a specific serialisation scheme to transmit information electronically, a syntax to make the content usable for machines, and some information model for semantic interoperability between the service provider and client.

An API is defined by a specification (the instructions for how it works), implemented by some software, deployed on a server accessible to target clients, and accessed through an addressable endpoint using some communication protocol. The content offered using the API is not necessarily part of the API definition, but an API might be specific to a particular kind of content (e.g. photographs, digital maps, time-series data, tabular data, weather data, sound recordings). In general the particular content offered is specific to a service endpoint.

6.3.2.1. Matching data and applications

In order to link data and applications, metadata for the data and for the application software must have sufficient information about the distribution protocol and format to determine whether an application can work with an offered distribution. The pattern is similar to that used by desktop computers to match files with applications, but more granular categorization of the distribution encoding format and information model is necessary to match data to applications offering functionality specific to that data.

From the software (data user) side of the connection, several factors come into play. The software might execute on an operating system like Linux, Windows, iOS, or Android, referred to here as 'standalone environment', or an application might execute in an online environment mediated by a Web browser application, referred to here as a 'browser environment'. In the standalone environment, an application has access to a local file system and various communication protocols and interfaces supported by the hosting operating system. We assume these include access to online resources using the internet. An internet-connected standalone application might implement a search function to find resources on the internet that work with the application.

In the browser environment, an application might run on a remote server and simply use the client browser as the user interface; at the other end of the spectrum, the server might download application code to the browser and the application runs in a browser container on the client side. In either case, the application execution would be started by an HTTP request to some Web location, and the HTTP protocol provides various mechanisms with the request that can be used to pass information specifying the target data for the application to open.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Search applications running in a Web browser will typically not be able to invoke an application to run outside of the browser's 'sandbox' for security reasons. In such cases, search results can list applications that will work with a particular dataset, and provide links to Web pages for additional information about how to install the application (if the user does not already have it installed), but the user will have to start that application and access the discovered data manually.

6.3.2.2. Data distribution (Data Provider conventions)

File download distribution. In the simplest and still common file-based data access scenario, the dataset distribution information in a metadata record includes a link (URL) that will get a file containing the actual resource content in a particular format. The format and information model for the file content must be specified in the distribution object in the metadata. The HTTP protocol is used to GET a resource. Given the view that an API specifies the functions offered and constrains the content of messages transmitted between a client and the agent, this simple file download is not considered an API.

Service based distribution. An API builds on a basic communication protocol (e.g. HTTP) by defining functionality and formatting to enable providing the specific data a user requires. This might involve filtering, subsetting, or various transformations for e.g. schema mapping, aggregating or anonymizing data. The focus here is on Web APIs that provide data using a URL for the endpoint location (the server that implements the data access protocol), with parameters to specify the particular data requested. The query parameters might be appended to this base URL as part of the URL, or provided as a message with the request. Metadata content requirements:

- **Service type** -- specify the kind of service. Ideally this should be a resolvable identifier. Currently there is no widely adopted registry for serviceType identifiers, in large part because services might be defined at different levels of granularity, and classifications might focus on function, data formats, thematic content, security, or other aspects of the service definition. For interoperability, there must be an external arrangement between data providers and consumers on the strings that will be used to specify service types. Cardinality: 1..*.

- **Service description document** -- many service implementations include provision for a document that provides a machine-actionable description of a service instance. Examples include OpenAPI documents, OGC Capabilities documents. Software designed to utilise a particular service type will typically include functionality to parse such a description document and engage with the service endpoint. If such a document is available for the service instance providing the resource distribution, it should be included in the distribution metadata. Cardinality (0..1, conditional, required if it exists)

- **Endpoint URL** -- the base location identifier (URL) through which the service is accessed. If a service description document is provided, the location of the service endpoint would be the other piece of information needed; the endpoint location might be related to the location for the service description document and thus not required to be specified separately. Cardinality (0..1).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- **Access constraints** -- Description of access privileges required to use the API, e.g. registration, licensing, payments. Note that access constraints applying to any distribution of the resource should be specified in the access constraints for the resource description as a whole.

If a service description document is not available, some basic information about the API should be provided in the metadata. The operations offered by the service and the output formats (serialisation scheme and information model) are typically defined in the service specification, and would thus inhere in the service type identifier for clients that recognise the service type. These might be optional in the service type specification, with choices for what it offered specific to a particular endpoint, in which case they should be asserted in the metadata for the particular endpoint.

- **Operations** - specify the functions offered by the service endpoint
- **Output formats** - specify the output formats (serialisation scheme and information model) for service response documents.
- **URL template** - A template for an HTTP service request that indicates how to invoke the service. The URL template must follow the conventions specified in IETF RFC6570[91]. Parameters that are specified by the user when invoking the service are enclosed in curly brackets ('{ }'). Note that the URL template must be consistent with the specified service type.
- **Parameters** - some description of the URL parameters that can be specified in the service request URL. This should include a description of what the parameter does, what data type is used, and the cardinality for the parameter. If the parameter is populated from a controlled vocabulary, some specification of the allowed values should be provided, either as an enumerated list or a link to a vocabulary that will be recognised by users.

6.3.2.3. Metadata distribution (Metadata Provider conventions)

Metadata providers offering APIs to search metadata catalogues can be considered a special case because they play a 'middleware' role between resource providers and resource consumers. The only real difference is in the intention of the content offered by the API. The resource they offer is data that is about other data, but the distribution description fits into the above content model. The service type would need to indicate that the API is for discovering information about resources (potentially in some thematic scope). The operations would necessarily include a search operation. The output formats would be the metadata schemes (and optional profiles) offered for service responses, e.g. ISO19115-3 MCP profile, ISO19139 INSPIRE profile, schema.org CDIF profile, DCAT-AP. URL template parameters would include the various properties that are queryable.

6.3.2.4. Software API connection (Data User conventions)

To identify an API that an application can work with, metadata for the application must specify what formats are acceptable for input data, and the interface(s) used by the application to request input data in that

[91] https://datatracker.ietf.org/doc/html/rfc6570

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

format. The software input file format will be matched with the output formats and the implemented communication protocols will be matched with the service types offered by resource distributions to determine where interoperability is possible.

- **Input file format** -- indicate input formats understood by the application. For interoperability, the format must be specified using strings established by convention between providers and consumers. Use of generic MIME-Types is not sufficient in most cases as these are quite general and can result in data-API matches that don't actually work.

- **Interface specification.** A communication protocol defines how messages are transmitted between a client and server. The software metadata must specify the application interface(s) (API) that it implements to interact with services providing resources for operation of the software. The API specifies the communication protocol that is used and the formats for requests and expected responses that an application uses to access data from a server.

- **Execution environment.** The software environment needed for operation of the application, e.g. operating system, online/cloud.

For applications operating in a single desktop or local-area network environment, operating systems like Windows, AppleOS, or Linux offer various communication protocols, and applications use various bespoke drivers to implement connection and communication. The simplest case is file-based access using the standard operating system file-open dialogs, and the 'Interface specification' is simply 'local file system'. File-based data retrieval using a URL is similarly simple and 'Interface specification' is simply 'HTTP'. The encodingFormat is the critical information to match data sources with applications in these cases.

The data source might be a relational database like PostgreSQL, MySQL, or any of a variety of noSQL datastores like SOLR, HaDOOP, or MongoDb. The 'Interface specification' in this case only needs to indicate that the application has drivers necessary to acquire data from one of these data sources by identifying the data source by name. The encodingFormat used to transmit data between source and consumer is invisible to the user in this case, so the input file format is not required to be specified.

An application might access data via a WebAPI-- using interfaces communicating via the internet and based on HTTP operations (GET, POST, PUT, DELETE), or by tunnelling operation requests embedded in HTTP requests (e.g. OGC GetCapabilities). An application might depend on some particular operations or request parameters (e.g. file formats or profiles), in which case the application metadata 'Interface Specification' should be specific to these requirements. Alternatively, the software might operate with any data source that implements a particular interface (API). In this case the software metadata does not need to specify particular file formats or request parameters, these are built into the software and the interface definition. For a particular data access connection, the dataset distribution needs to specify the correct request parameters to get particular data (see **Service based distribution**, above).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.3.3. Publishing metadata

This section describes workflows recommended for a metadata publisher to make their metadata accessible to Web crawlers for search engines to index their resources. Figure 2 (below) is a flow chart showing the decision tree to determine how to expose metadata. Numbers in the following discussion refer to numbered boxes in the diagram.

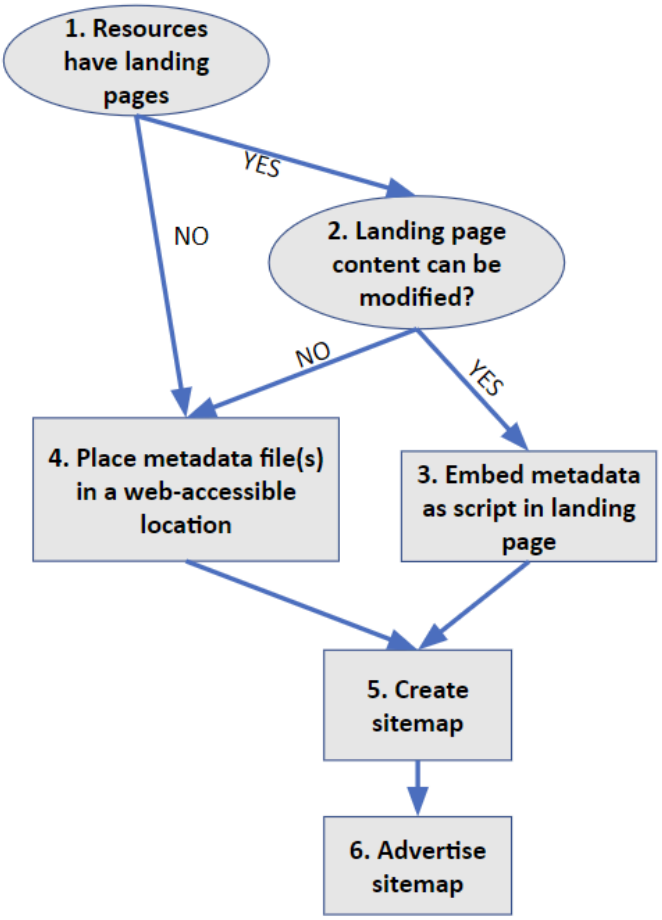

Figure 3. Decision graph to determine where metadata is located, from the resource publisher's perspective. Steps are numbers and referenced in text.

6.3.3.1. Option 1. Embedded in HTML

Starting at the top (1), if there are HTML landing pages that describe the resources of interest, and the metadata publisher has the necessary authority to update the content of these pages, then CDIF metadata serialised as JSON-LD (see Appendix 1) should be embedded in an HTML <script> element in the <head> section of each landing page (3). The script should have the following type and profile attributes:

type="application/ld+json" profile="CDIF1.0"

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.3.3.2. Option 2. Individual metadata file URLs

If the resources of interest do not have individual landing pages, or the metadata publisher does not have authority to update the content of landing pages, the metadata should be placed in a Web-accessible location (step 4 in Figure 2). There are two common approaches:

- Each metadata record is accessed in a separate, static file with its own URL. The CDIF metadata is serialised as JSON-LD (see Appendix 1). MIME type for the metadata file, returned as the Content-Type parameter in the HTTP response header, is:

  type="application/ld+json" profile="CDIF1.0"

- Each metadata record is accessed dynamically from the server using a URL. There are various open-source metadata server systems that can be configured to deliver CDIF metadata from the server's metadata database, e.g. GeoNetwork OpenSource[92], GeoPortal[93], CKAN[94]. The metadata retrieval URLs have different syntax depending on the software used, but typically include a metadata record identifier and a format parameter that would be used to indicate that CDIF metadata should be returned. If there is a format parameter in URL requests, its value should be 'CDIF1.0'.

6.3.3.3. Option 3. Metadata list file

- A collection of metadata records is gathered in one file accessed using a single URL. For CDIF, this file should contain a set of CDIF JSON-LD metadata objects, implemented as a schema.org ItemList[95]. See example in Appendix 1. The MIME type for the collection is:

  type="application/ld+json" profile="CDIF-list-1.0"

6.3.3.4. Find metadata

CDIF recommends the use of sitemaps to address the questions of how crawlers find metadata to index or use. A sitemap[96] is an XML document that a metadata crawler or harvester can access (6 in Figure 2). The basic sitemap is an XML document that is a structured list of URLs, with an optional date stamp property that should indicate when the metadata at the URL target location was last updated. The most basic CDIF recommended approach for resources that have landing pages that can be modified is Option 1 outlined above. The workflow for a harvester in this approach corresponds to existing structured data on the Web practice, with metadata conforming to the CDIF recommendations. Existing robots.txt and sitemap files can be used, and no modifications need to be made to HTTP headers provided by resource servers.

[92] https://geonetwork-opensource.org/
[93] https://github.com/Esri/geoportal-server-catalog
[94] https://ckan.org/
[95] https://schema.org/ItemList
[96] https://www.sitemaps.org/protocol.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

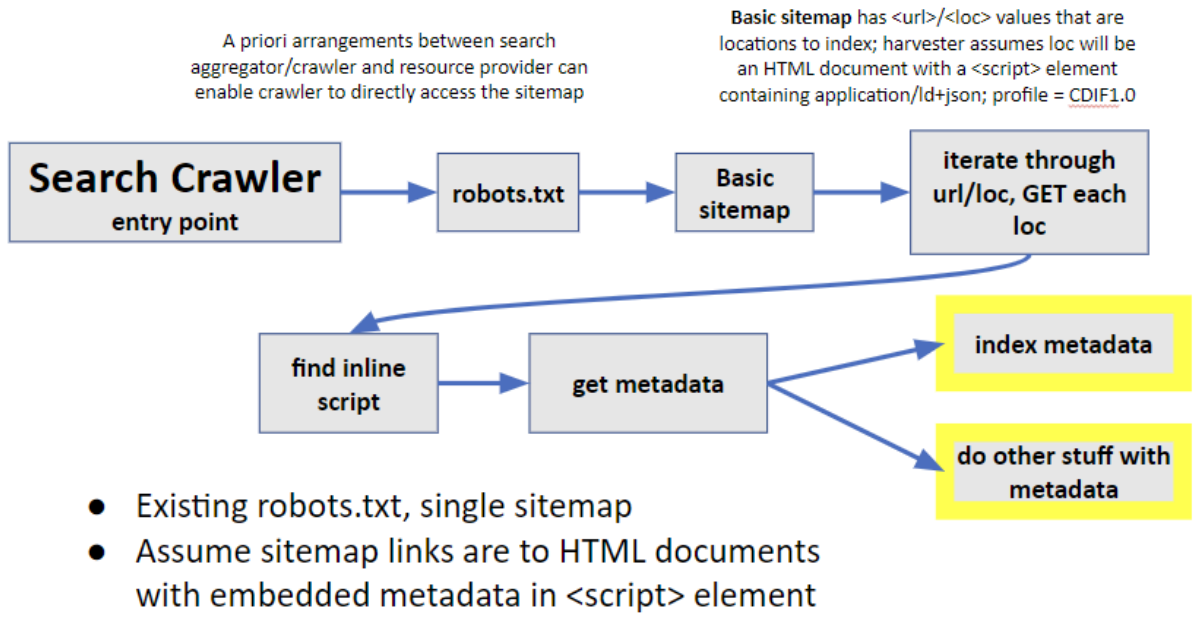

Figure 4. Basic harvesting pattern: sitemap with locations that contain HTML documents with embedded CDIF metadata

If resources do not have landing pages, or the landing pages cannot be modified, harvesters must be provided with links to the metadata records to index. The basic approach for this next level of service is similar, but the harvester cannot assume that the metadata is embedded in the content accessed from the sitemap url/loc. The sitemap url/locs must either get the metadata document directly, or it might get the resource content directly. In the second case, a URL to get the metadata must be included in the HTTP response header. Because the basic sitemap only provides a URL, the harvester will need to check which of these options is being used. The approach using standard Web architecture is to inspect the HTTP response header. If the returned document is a CDIF metadata document, the header will have a content-type parameter with the value "application/ld+json; profile='CDIF1.0'" or profile='CDIF-list-1.0' if the document is an itemList with multiple metadata records. The harvester should get the content at the sitemap url/loc and use that. If the content-type has a different value, then the header should have link element with rel='describedby', type='application/ld+json', and profile='CDIF1.0' or profile='CDIF-list-1.0'. In this case, the harvester will GET the content at the link href and use that. Inclusion of the describedby link in the HTTP header is the pattern used by Signposting.

Other Signposting links (see Appendix A. 3) could be inserted into the HTTP response headers (or HTTP landing pages if they exist) if Web site administrators allow. This can provide additional value for clients implementing signposting.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

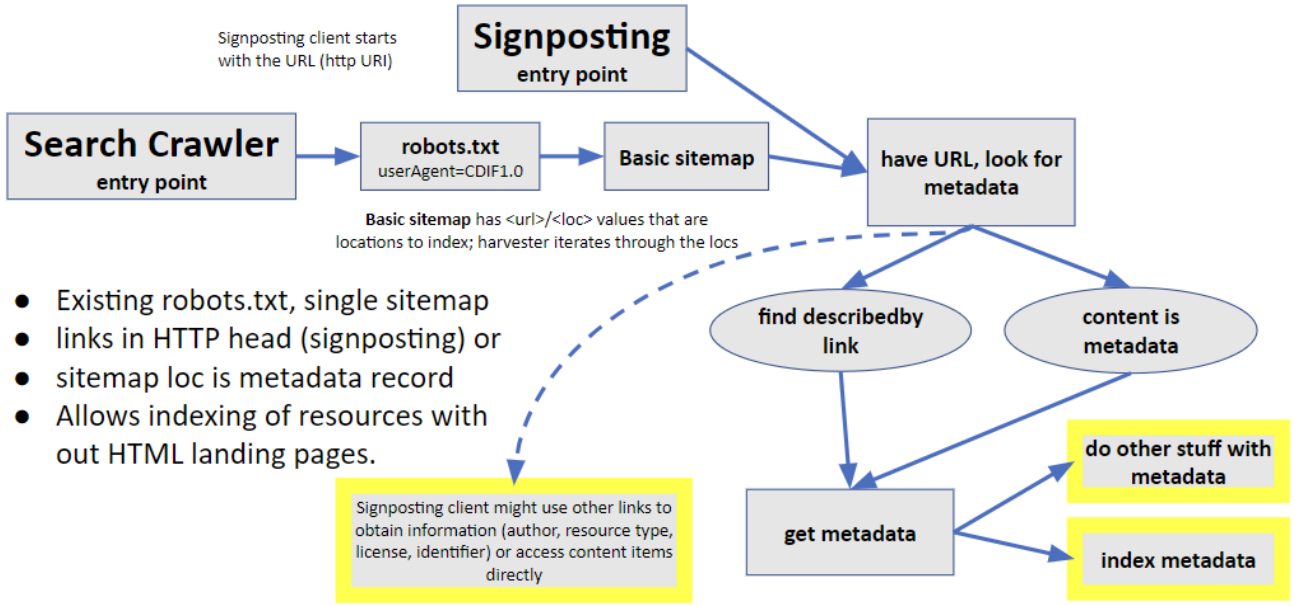

Figure 5. Harvesting pattern for locations that do not have embedded metadata scripts. Either the url/loc content is a metadata document (JSON-LD, CDIF profile), or there is a Signposting link with rel=describedby in the HTTP head. To indicate that this is not a standard sitemap, the robots.txt link to the sitemap should identify CDIF1.0 as the userAgent.

Note that in this second case, the sitemap url/loc locations will get content that is not HTML. A non-CDIF aware harvester will be expecting HTML (or indexable text) content, and this might cause problems. Harvesters will generally find the sitemap using a link from a robots.txt[97] file placed in the root of the server containing the sitemap and metadata. In the robots file, the user agent value can be used to indicate a sitemap link for CDIF-aware agents in cases for which the sitemap links are not to pages that have embedded CDIF metadata <script> elements. If the harvester accesses a sitemap under this user agent, the url/loc location content should be processed as outlined in the last paragraph.

```
User-agent: CDIF1.0

Sitemap: http://www.example.com/CDIFsitemap.xml
```

Based on these recommendations, metadata generated using the CDIF content and serialisation can be found and harvested by agents using standard off-the-shelf Web technology.

6.3.4. The Data Catalog Vocabulary (DCAT)

While the discussion up to this point has focused on the use of Schema.org, it is recognised that DCAT is also an important specification for exposing discovery metadata. To support those who use DCAT, we recommend that the DCAT metadata be expressed in JSON-LD as recommended for Schema.org, and the equivalent fields be used.

[97] https://datatracker.ietf.org/doc/rfc9309/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

DCAT has become very common in some areas: DCAT-AP and GeoDCAT have gained a lot of traction in Europe. DCAT-US promises to gain similar traction. While this varies domain-by-domain, the fact that DCAT itself is designed for describing data, while Schema.org is immensely broad in scope (it describes basically everything for the purposes of the Web!), means that DCAT has some features which are specifically useful for CDIF's purposes.

There are several different mappings between Schema.org and DCAT available, and some may be more appropriate than others for any particular implementation. By default, however, we recommend the mapping from the W3C group which has developed DCAT:

1. **DCAT - Schema.org mapping in the context of DCAT version 2:** https://ec-jrc.github.io/dcat-ap-to-schema-org/ and also https://www.w3.org/TR/vocab-dcat-2/#dcat-sdo with https://w3c.github.io/dxwg/dcat/rdf/dcat-schema.ttl being a mapping between DCAT2 and SDO 3.4 . "...This mapping is axiomatized using the predicates `rdfs:subClassOf`, `rdfs:subPropertyOf`, `owl:equivalentClass`, `owl:equivalentProperty`, `skos:closeMatch`, and also using the annotation properties `sdo:domainIncludes` and `sdo:rangeIncludes` to match [SCHEMA-ORG] semantics. The alignment is summarised in the table below, considering the prefix `sdo` as `http://schema.org/`. …" .
2. **Update in the DCAT version 3 documentation:** https://www.w3.org/TR/vocab-dcat-3/#dcat-sdo

6.3.5. Resources for this profile

Cox S.J.D., Gonzalez-Beltran A.N., Magagna B., Marinescu M.-C., 2021, Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17(6): e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Weigel, T., Plale, B., Parsons, M., Zhou, G., Luo, Y., Schwardmann, U., Quick, R., Hellström, M., Kurakawa, K. (2018). RDA Recommendation on PID Kernel Information (Version 1). DOI: https://doi.org/10.15497/RDA00031

6.4. Data access

6.4.1. Background

6.4.1.1. Problem statement

We make a distinction between 'access' and 'usage'. The first relates to the activities which mediate the initial retrieval of a digital object for research; the second relates to all subsequent operations performed on

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

that digital object. The definition and description of these activities are bounded by access and usage policies.

These access and usage policies (herein 'access policies') are, if they exist at all, unstructured and bespoke. Data providers may not make access policies explicit and when they do, they tend to re-invent new policies locally. Therefore data users experience new data access policy content and structure at every access-related interaction across the science system. Any kind of aggregation or orchestration of the data across providers is stymied by an incoherent data access policy environment in terms of existence, coverage, content, and machine-actionability.

6.4.1.2. Objective

The objective of including 'access' in the Cross Domain Interoperability Framework (CDIF) is to progress from the current unpredictable, cottage industry, free-text based descriptions of access policies to a more structured and standardised, machine-actionable approach. The benefits and convenience for data requestors accrue through:

- transparency and efficiency in requesting data;
- consistency of access experience across data custodians;
- automated processes and services across those providers;
- federated permissions query and access.

For data custodians, the structured and standardised approach to data asset permissions through a structured ontology (e.g., ODRL) provides:

- Default minimum good practice guidance and design patterns for new systems development;
- Efficiencies, equity, and transparency in processing access requests;
- Ability to track and evaluate access requests to improve processes and address inequities in access and reuse;
- Opportunities for repository or archive providers to participate in multi-organisational collaborations to support collaborative science.

6.4.1.3. High-level recommendation

To promote interoperability and mutual intelligibility around access conditions, CDIF recommends use of the Open Digital Rights Language (ODRL) to describe data asset access policies. ODRL is a RDF-based, "widely adopted language for expressing permissions, obligations, and conditions related to digital rights"[98] that can be serialised in JSON and XML.

While minimal in terms of classes and relationships, it nonetheless allows sufficient flexibility through the use of constraints and refinements and descriptive logic. ODRL also allows straightforward extensions to the core model and vocabulary with ODRL Profiles.

[98] Policy Patterns for Usage Control in Data Spaces 20 Sep 2023, https://arxiv.org/pdf/2309.11289.pdf

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The core classes are:

- **ODRL "Parties**" - actors or agents who exercise:
- **ODRL "Rules"** - ODRL Permissions (can do), ODRL Duties (must do), ODRL Prohibitions (must not do) encompassing:
- **ODRL "Actions"** - operations performed on:
- **ODRL "Assets"** - digital objects in a CDIF context.
- **ODRL "Policies"** are structured artefacts which encapsulate the above information.

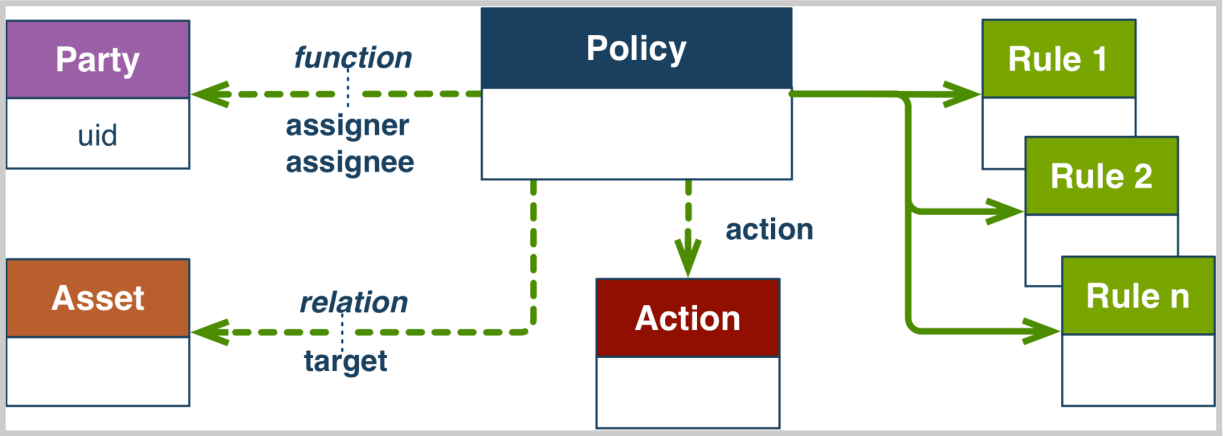

Figure 6. Simplified schematic of main ODRL Classes.

6.4.1.4. Risks and enablers for CDIF using ODRL

The existence of ODRL as W3C standard is itself the key 'enabler' for CDIF, since CDIF is committed to using existing, well-supported standards to support interoperable practice wherever possible.

The first risk with ODRL is that (as with all standards) it doesn't cover all scenarios. The scenario-based discussion below aims to tease out what valuable work can be done with ODRL immediately and which scenarios might need further extensions or new approaches. Our conclusion is that there is enough to get on with using ODRL as is, as well as enough extension/profiling capability in ODRL. This risk of insufficient coverage is therefore low in the first instance and well mitigated in the second.

The second risk is the barrier to adoption of ODRL. There is an urgent need for tooling to simplify how domain scientists and data curators might interface with this mature W3C standard. The information model, ontology, and linked data implementation are beyond the capability of many of the scientists and infrastructure service providers who are on the 'critical path' information pipeline for creation of these standardised access policies at scale. This risk is both high and likely and currently has no mitigation.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

A Google search will locate a small number of ODRL snippets that can act as basic templates but without a reasonable degree of familiarity with RDFS and OWL, it is not straightforward to construct the more complex ODRL policies that will be required to model and describe real-world access processes.

6.4.2. Scope

This recommendation is about the use of ODRL for interoperable, machine-actionable, expression of data access policies. It is important to note this is not the entirety of 'Access' from a plain English or FAIR perspective. Accessing sensitive data is a complicated, multi-faceted, multi-party process involving for example:

- Clarifying intellectual property rights, and, where appropriate, copyright.
- Negotiation with data producers
- Specifying data reuse licences
- Establishing bespoke legal agreements
- Ensuring cybersecurity
- Assessing data sensitivity and identifying potential secondary disclosure risks
- Preventing reverse engineering of the data
- Establishing the ethical and legal constraints pertaining to the nature of the data in question
- Clarifying the nature of a specific access, including the requestor and the data request purpose
- Identifying roles or attributes upon which access may be granted
- Provision of secure access environments with or without analytics

This recommendation acknowledges the importance of all those (and related) elements but focuses on standard ways of capturing data access requests in structured, machine-actionable policy statements at the point of access, as well as prescribing what subsequent operations can (or cannot) be performed until the end of the research lifecycle. For example this paper does not offer guidance on how to classify data sensitivity. It does provide a recommendation on how to express access constraints based on a data provider's pre-existing data sensitivity classification. It doesn't explain when, why or how an ethics approval is needed. It does provide a recommendation for the standard machine-readable policy that would tell a data requester that an ethics approval is needed in order to retrieve the data.

In scope for this recommendation are:

- The standard expression of access policies
- The "publishing" of those policies
- The automated execution of machine-readable access policies.

Useful enablers of an access policy but 'out of scope' for this recommendation are:

- How access policies are derived
- How to classify of the sensitive nature of the data
- How to assess the risk of secondary disclosure

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- Licences / Legal Agreements /IP / Copyright
- Good practice IT Security/ cybersecurity
- Consent/ Ethics Approval.

6.4.2.1. Data provider vs. data user access policies

ODRL data access policies can be formulated or executed by either the data provider or the data user. For example, a data user may establish an ODRL policy that requires the data to conform to a certain standard as part of a machine readable request. Because of the audience and context of CDIF, this document starts from the perspective of the data policies of the data provider, since they fit in with the perspective of a metadata standard for interoperability being established for/by data providers.

6.4.2.2. Access classifications

It is not uncommon for data providers (or communities thereof) to apply broad 'access classifications' to data, for example:

- openAccess / restrictedAccess[99]
- Open / Safeguarded / Controlled[100]
- ClosedAccess / EmbargoedAccess / RestrictedAccess / OpenAccess[101]

When included in discovery metadata, such classifications drive useful functionality in portals or catalogues for filtering or faceting search results. They are not sufficient however to drive automatable access processes. Machine-actionable access policies are inherently more fine grained than such high level classifications and involve a number of top level entities (parties/actions/rules) over sequenced workflows (request/access/usage). It would not be practical to create a top-down classification system granular enough and with broad enough consensus to drive machine actionable access policies. Rather the recommendation here is to use a standardised and structured language for access policies so that these can be generated bottom-up and yet retain the appropriate level of interoperability.

6.4.3. Access policy scenarios and recommendations

This section analyses a spectrum of scenarios that identify typical drivers and needs for interoperability of data access policies:

1. Automating internal access processes
2. Aggregating access policy metadata across a community
3. Shared services for request and fulfilment
4. Federated analysis over multiple sensitive data providers

[99] CESSDA https://www.cessda.eu/
[100] UKDS https://ukdataservice.ac.uk/
[101] OpenAIRE https://www.openaire.eu/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

These scenarios build a set of typical requirements for the interoperability of access policy summarised in the figure below and detailed in the following sections. Each section starts to formulate recommendations for CDIF adherents to use ODRL and other standards to achieve their interoperability needs.

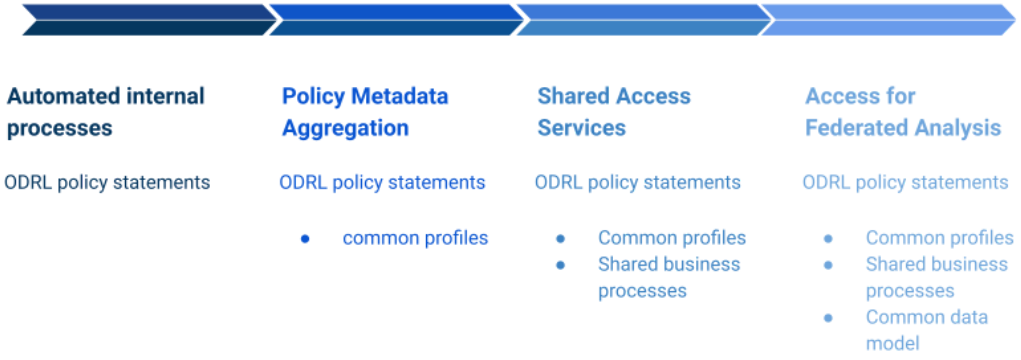

Figure 7. Data access scenarios

6.4.3.1. Scenario 1 - Standalone repository (dataset download)

Summary

A standalone repository can hold sensitive personal data requiring mediated access. The target CDIF persona here might be a data manager in a high profile research group in a medical research institute that conducts health studies. The organisation is committed to providing 'secondary access' to other research groups to build their own collaborative network and to optimise the reuse and impact of data they have collected in the course of their research.

This data asset manager is looking to make the access processes visible, to develop access policies to ensure consistent, efficient processing of access requests, and to provide standardised categories that could support some simple automation of local access workflows. The group is looking to make a start with one overarching policy for its clinical trial datasets that makes it clear that these datasets are available for secondary use by other academic users. The group also wishes to specify the actions that can be performed on those assets by end users (e.g. access for academic researchers for a specific research question to the full dataset in a trusted research environment [TRE]).

ODRL Recommendations

In this scenario, the data custodians would use ODRL to express the data asset access policy by first decomposing the access conditions, actors, and actions into:

- Registered academic researchers, who can demonstrate accreditation by ANU, can analyse dataset 6123 in a TRE environment to address TB drug efficacy for a defined population

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- Pseudonymized data available for secondary use by any registered researcher based in Australia.

Example ODRL Policy #1

```
{
  "@context": "http://www.w3.org/ns/odrl.jsonld",
  "@type": "Set",
  "uid": "http://example.com/policy:1",
  "permission": [
    {
      "target": "http://example.com/dataset_6123",
      "action": "Analyse_In_TRE",
      "assignee": {
        "@type": "PartyCollection",
        "source": "ANU_Registered_Users",
        "refinement": [
          {
            "leftOperand": "Accredited_Researcher_Property",
            "operator": "eq",
            "rightOperand": "true"
          }
        ]
      }
    },
    {
      "target": "http://example.com/dataset_6123",
      "action": "SecondaryUse",
      "assignee": {
        "@type": "PartyCollection",
        "source": "ANU_Registered_Users",
        "refinement": [
          {
            "leftOperand": "Researcher_Location_Property",
            "operator": "eq",
            "rightOperand": "Australia"
          }
        ]
      }
    }
  ]
}
```

NB: In the example above, **Actions**, **PartyCollections** and **Refinements,** while now framed in an ODRL structure, only have local meaning and can only be executed locally. That is to say, 'Analyse_In_TRE' and 'SecondaryUse' are labels that currently have no wider potential for machine-actionability beyond the local repository environment.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.4.3.2. Scenario 2 - Access metadata aggregation

Summary

Rather than data users interacting with individual data assets, data users may want to query a catalogue of data assets to understand which assets may be used to address their specific data reuse needs.

The CESSDA Catalogue[102] is an aggregated social science data catalogue which is populated with metadata from 19 'Service Providers' (SPs) spread across Europe. These SPs make discovery metadata available for harvesting by CESSDA, predominantly through either an OAI provider[103], a Dataverse instance[104], or a Colectica instance[105]. Discovery metadata must be provided in either DDI Codebook[106] or DDI Lifecycle[107] and its consistency is assured by the use of DDI Profiles[108].

In CESSDA, metadata relating to access is currently framed in prose and while this prose may be encapsulated formally in an XML element, e.g., dcterms:accessRights or ddi:accessRights, this does not supply the data consumer with specific, structured information about how they should access data and under what conditions. The practical consequence of this is that a researcher will be directed to another SP's catalogue from CESSDA, only to then find that a dataset requires a formal application process, or that it may not even be available to them as a citizen of another country.

Formalising access descriptions with a common set of semantics and syntax such as ODRL would enable a discovery metadata aggregator, such as the CESSDA Catalogue, to aggregate content about access, making access conditions transparent *at the point of aggregation*. There is currently no filter on the CESSDA Catalogue for 'access conditions' and for good reason i.e. the heterogeneity of the access descriptions that are supplied by multiple SPs. One posited approach is to overlay this heterogeneity with two top-level labels 'Open', and 'Restricted' cf. CESSDA Data Access Policy[109]. This merely organises a large set of unstructured prose items into two groups of unstructured prose items where the border between 'Open' and 'Restricted' is fuzzy.

ODRL recommendations

One typical access scenario for CESSDA repositories involves making datasets available on the basis of a researcher's country location before the researcher is permitted to download the data. In the CESSDA catalogue, there is currently no simple way to identify datasets which fall under this scenario. The best we

[102] https://datacatalogue.cessda.eu/
[103] https://www.openarchives.org/pmh/
[104] https://dataverse.harvard.edu/
[105] https://www.colectica.com/software/repository/
[106] https://ddialliance.org/Specification/DDI-Codebook/2.5/
[107] https://ddialliance.org/Specification/DDI-Lifecycle/3.2/
[108] https://github.com/cessda/cessda.metadata.profiles
[109] https://zenodo.org/record/6722000

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

can do is to drill down into the 'Restricted' or 'Open' filters and read the specific licence/rights/access prose for each individual dataset.

Using a common ODRL policy to express the above access conditions, we can begin to support more granular groupings of datasets based on their common access features. Over time, a set of re-usable ODRL policies could be deployed across multiple repositories. At a minimum, this would require referencing a number of standard items such as ODRL 'Parties' or ODRL 'Actions' by URI.

ODRL Policy #2a - potential for re-usable 'Parties' and 'Actions'

```
…
"permission": [
      {
        "target": "http://example.com/example_digital_object",
        "action": "See Note 1",
        "assignee": {
          "@type": "PartyCollection",
          "source": "See Note 2",
          "refinement": [
            {
              "leftOperand": "See Note 2",
              "operator": "See Note 2",
              "rightOperand": "See Note 2"
            }
          ]
        }
      }
]
…
```

NB

1. A typology of 49 common, reusable Actions are defined in the ODRL Core Vocabulary[110]. Surprisingly, this doesn't include the ubiquitous 'Download'. Such additional terms can be defined in 'ODRL Profiles' (outside of the scope of this document). ODRL Profiles are an extension mechanism for the core model and vocabulary.
2. Real-world access policies will typically define which types of users can, cannot or must perform a particular action. Within the constraints imposed by ODRL syntax, the approach taken above is to assert a generic 'PartyCollection' - essentially the universe of all possible agents - and then refine it to a particular subset of agents with required attributes. In practice, there is a finite number of 'refinement' permutations that could be pre-defined and assigned a persistent URI to be used as the 'source' key above alongside the dereferenced {leftOperand, operator, rightOperand} array.

ODRL Policy #2b - referencing ODRL 'Parties' and 'Actions' by URI

[110] https://www.w3.org/TR/odrl-vocab/#actionsCommon

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
…
"permission": [
      {
        "target": "http://example.com/example_digital_object",
        "action": "http://cdif.org/odrl/1/download",
        "assignee": {
          "@type": "PartyCollection",
          "source": "https://cessda.eu/partycollection:312",
          "refinement": [
            {
              "leftOperand": "http://www.w3.org/ns/odrl/2/spatial",
              "operator": "http://www.w3.org/ns/odrl/2/eq",
              "rightOperand": "de"
            }
          ]
        }
      }
]
…
```

NB: This policy asserts that users located in Germany have the 'Download' permission. The reusable 'PartyCollection' item https://cessda.eu/partycollection:312 (signifying users located in Germany) is re-used from an existing vocabulary used by multiple repositories.

6.4.3.3. Scenario 3 - Shared service provider for requests and fulfilment (clearing house)

Summary

A business driver for interoperability can be the desire to 'componentize' services or functions and provide them once in a consolidated/specialised way across an ecosystem. For example a government or funding council might calculate that they are resourcing clerical staff and technology services in multiple data providers to field and fulfil data requests that emerge from a single national pool of researchers. Creating a shared service or common clearing house would be one response from the point of view of the central funding agency. Benefits of a clearing house approach could include:

- Streamlining processes for data users
- Support for combined interdependent requests across multiple data providers for complex studies
- Efficiencies of resourcing and specialisation
- Natural combined reporting and tracking of usage and impact

An example of this kind of thinking is the Australian Government's DataPlace[111], "a new, whole of government digital platform that will make it easier to request Australian Government data." The trend to

[111] https://www.dataplace.gov.au/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

replace siloed departmental government services with “whole of government” platforms is a result of the combination of drivers for efficiency with the new opportunities that information systems and technology provide.

Standardised, machine-actionable data access policies are one such interoperability technology advancement that help to deliver on two assumptions of a shared clearing house:

- that the cost of sharing the information about requests does not overwhelm the savings in staff and systems.
- That a central capability can actually action things and lighten the load of the network rather than just re-routing jobs for fulfilment back to the participating node for ‘local implementation’.

A common language for expressing data access policies across the cooperative of data providers using the clearing house is a necessary but insufficient first step. A shared understanding of how policies interface with business processes is also required so that both data provider and the clearing house can be satisfied that request fulfilment was efficient, consistent and auditable.

ODRL recommendations

The schematic below illustrates a notional message exchange initiated by a researcher requesting a dataset stored in a specific repository with a central ‘clearing house’ or brokerage mediating the access on behalf of this and many other repositories. Note that here, the Broker is not acting as a portal for finding the data, but as a mediator of access to a resource known to be at the Repository.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

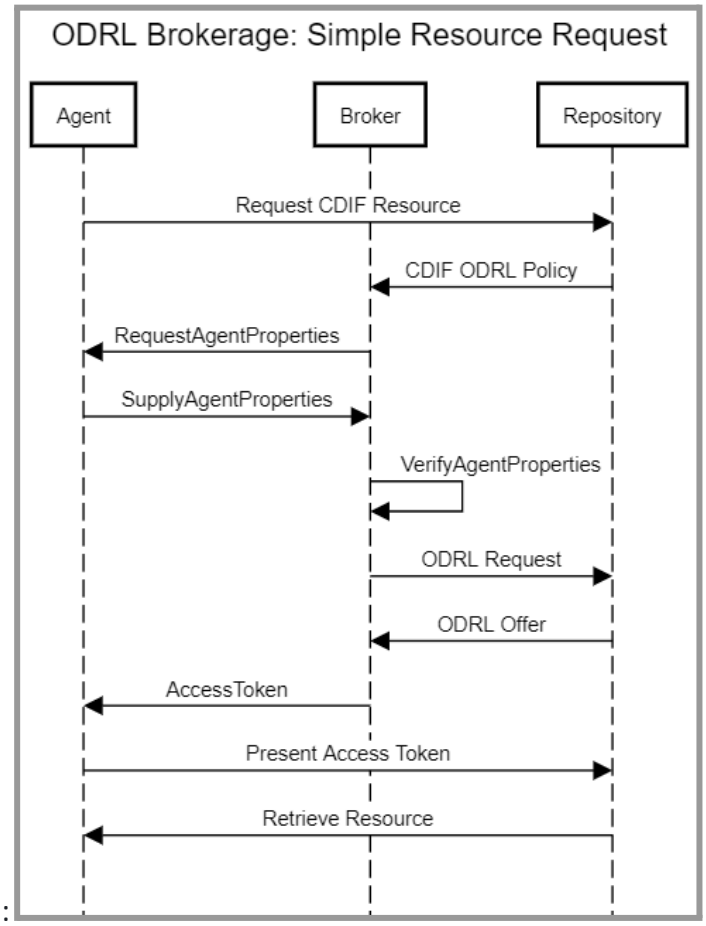

Figure 8. ODRL brokerage of simple access request

Message flow description:

#	Message	Message pseudo-description
1	Request CDIF Resource	User locates dataset resource on repository catalogue with DOI https://dx.doi.org:12345 and clicks on 'Access Data' button
2	CDIF ODRL Policy	Repository posts new RequestFulfilment object to Brokerage API including location of CDIF record for https://dx.doi.org:12345, a JSON-LD artefact which encapsulates ODRL Policy #2b
3	Request AgentProperties	Brokerage parses ODRL Policy #2b and determines that the user must be located in Germany. This is a very reductive example but more complex real-world

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

		examples might be the collection of additional information through an online form managed by the broker.
4	Supply AgentProperties	The user will supply any information requested by the broker.
5	Verify AgentProperties	The brokerage will perform the necessary validation and verification of any information supplied by the user. This might involve auxiliary processes, such as seeking third party approval.

6	ODRL Request	Once all conditions outlined by the ODRL Policy have been met and satisfied by the broker, an ODRL Request will be sent to the repository on behalf of the user.
		
		{
		"@context": "http://www.w3.org/ns/odrl.jsonld",
		"@type": "Request",
		"uid":
		"http://brokerage.com/odrlrequest/9a112bc6-d93d-4a59-8384-0ac65399ef94",
		"permission": [
		{
		"target": "http://repository7987.org/dataset_6123",
		"action": "http://cdif.org/odrl/1/download",
		"assignee": "http://orciduser675765",
		}
		]
		}

7	ODRL Offer	The repository will respond with an ODRL Offer to the broker.
		
		{
		"@context": "http://www.w3.org/ns/odrl.jsonld",
		"@type": "Offer",
		"uid":
		"http://brokerage.com/odrloffer/2f93930b-93f7-418f-a045-aa49d09faf2b",
		"permission": [
		{
		"target": "http://repository7987.org/dataset_6123",
		"action": "http://cdif.org/odrl/1/download",
		"assigner": "http://repository7987.org"
		}
		]
		}

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

8	AccessToken and ODRL Agreement	Assuming the broker accepts the offer on behalf of the user, an encrypted Access token is sent to the user and an ODRL Agreement is filed for audit purposes.
		
		{
		"@context": "http://www.w3.org/ns/odrl.jsonld",
		"@type": "Agreement",
		"uid":
		http://example.com/odrlagreement/d91d4663-f9bd-4bd5-8a81-c331e19bf987",
		"dcterms:references":
		[
		
		"http://brokerage.com/odrlrequest/9a112bc6-d93d-4a59-8384-0ac65399ef94",
		"http://brokerage.com/odrloffer/2f93930b-93f7-418f-a045-aa49d09faf2b"
		],
		
		"permission": [
		{
		"target": "http://repository7987.org/dataset_6123",
		"action": "http://cdif.org/odrl/1/download",
		"assigner": "http://repository7987.org"
		"assignee": "http://orciduser675765",
		}
		]
		}

9	Present Access Token	The user presents the token to the Repository
10	Retrieve Resource	The repository makes the file available for immediate download to the user.

Now, this is indisputably over-engineered for a simple download where the only condition, as per ODRL Policy 2b (above), is that the researcher be based in Germany, However, in the wild, access policies are multi-step processes with complex criteria, often with asynchronous handoffs to third parties for approval. The framework above, while necessarily brief for the purposes of this document, illustrates how a more complex stateful negotiation process could be performed using ODRL as the basis for defining conditions, then executing those conditions through a centralised brokerage, then generating a rich set of accounting information for access requests & fulfilment across any group of repositories where economies of scale would be desirable.

From defining ODRL Actions to executing ODRL Actions

Interoperability is sometimes referred to as a stack with building blocks that start with technology, but then need successive layers of semantic, business process, governance and legal interoperability to achieve

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

effective inter-operation. Scenario 3 in this recommendation takes the CDIF requirement up into the "process interoperability" level.

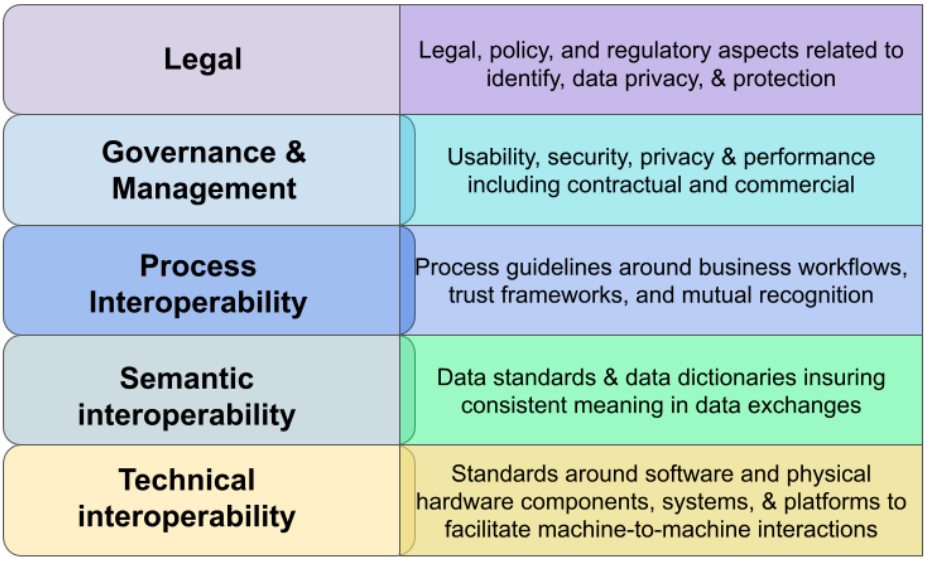

Figure 9. Interoperability framework - 5 building blocks. Adapted from the European Commission's JoinUp initiative [112]

It should be noted that many different communities have developed a similar view of a levelled interoperability. In the official statistics community, we see a very similar view in the UNECE Data Governance Framework for Statistical Interoperability (DAFI).[113]

The ODRL examples in previous sections have introduced the principle of disambiguating prose components of an access policy by using URIs as references to (i) pre-defined process actors (ORDL 'Parties') and (ii) pre-defined procedural steps (ODRL 'Actions'). By leveraging the ODRL model, these are framed in consistent syntactical ways with less ambiguous semantics.

While these would constitute a valuable move towards (A)ccessibility and (I)nteroperability in the loose FAIR sense, these examples do not in themselves provide sufficient detail or semantics for the automated implementation of ODRL Actions across multiple parties. That is to say, what has been described so far is more about machine-readability than actual machine-actionability.

The ODRL Common Vocabulary defines 49 Actions[114] which are operations that may be done (ODRL Permissions), or must be done (Duties) or must not be done (ODRL Prohibitions).

[112] https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers
[113] https://unece.org/sites/default/files/2024-03/HLG2023%20DAFI%20Final_0.pdf
[114] https://www.w3.org/TR/odrl-vocab/#actionsCommon

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

ODRL stops there. To enable repeatable and consistent implementation of request fulfillments framed by ODRL policies, we will need to identify how ODRL invokes machine-actionable workflows, an example being Common Workflow Language (CWL)[115] as well as what downstream invocations might be necessary, such as creating on-demand secure analytics environments (Infrastructure-as-code or IaaC).

Practically, taking CDIF to the next level of maturity in the access arena will mean committing to the creation of a CDIF ODRL Profile that extends the core ODRL ontology and vocabulary so that a widely applicable set of cross-domain Actions are available for re-use as well as being associated with unambiguous workflow implementations, expressed with additional classes and attributes in the CDIF ODRL Profile.

6.4.3.4. Scenario 4 - Federated analytics

Summary

Federated learning, also called distributed or multiparty learning or computation, distributed analysis, or distributed learning, is a way to reuse data through code without moving the data from the place where data are collected or stored. In contrast to other approaches to dealing with sensitive data where the data are coarsened through anonymisation to protect data subjects, federated learning allows for the reuse of sensitive data in its original format. Beyond ensuring a privacy-by-design approach through multiparty computation and differential privacy to prevent the reverse engineering of individual data centres, sets, and subjects, federated learning is a useful approach when the data are too large to move and supports more efficient (i.e., greener) model estimation. Federated learning is predicated on interoperable, high quality data described by interoperable, high quality, rich metadata.

Federated learning facilitates the real-time use of a living dataset, in contrast to approaches where a time-fixed version of a dataset is submitted to a repository or trusted research environment and subsequently shared. As such, federated learning is an important tool for enabling rapid reuse of time-varying sensitive data, as in the health data space. Federated learning is used for reusing participant-level data from hospitals, e.g., FAIR Data Point initiative in the Netherlands[116], where federated approaches ensure sensitive, participant-level data can be reused across hospitals even though data can not be transferred from one hospital to another because of local data protection legislation. Federated learning is most often applied when analysing big, sensitive data, as in remote sensing applications that reuse sensitive satellite imaging data.

Federated learning approaches generally have two levels of permissions, one at the level of the research question and initiative, i.e., this specific group of data assets will participate in a given initiative to answer a given research question, and one that reflects access conditions at the individual asset level. Although humans are engaged in providing permission for data assets to participate in federated learning, federated

[115] https://www.commonwl.org/
[116] https://catalog.accesshealthdata.nl/#/home

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

learning is dependent on machine-to-machine communication of the types of actions possible for the data assets that are queried in the federated learning approach.

In a federated analysis approach, the data user is the software agent. The agent is not so interested in 'discovering' access and usage policy. These have been set up at the establishment of the research initiative. Machine-actionable access policies are here used during a machine-to-machine transaction to verify the 'identity' of the agent, allow access to the target data asset(s) or parts thereof, and check that the actions performed are those that are allowed.

ODRL recommendations

Specify ODRL Action

This Scenario underlines the need for specifying ODRL 'Actions' relevant to federated learning or other analytical actions. The current ODRL 2.2 list of Actions for Rules does not include any Actions relevant to analytics. As a representative example, an ORDL policy guiding, controlling, constraining distributed analytics might want to oblige the software Agent to only perform a specific type of Action, say 'apply this pre-agreed mathematical process'.

It is recommended that ODRL be used for these use cases, but an extended typology of Actions be developed for distributed analytics (as part of a CDIF profile).

Invoke workflow with ODRL Actions

Federated learning of sensitive data needs to be accompanied by machine-actionable permissions and common work flows to allow machines to appropriately apply actions to data assets given the rules set for the federated analysis and the specific interactions with the individual data assets contributing to the federated analysis. As such, ODRL coupled with an structured, machine-readable approach to workflow execution is a necessary part of federated approaches to data reuse.

Therefore this Scenario shares many of the ODRL requirements of the previous scenario around integration of ODRL elements with a business process. Here again we would need to 'identify how ODRL invokes machine-actionable workflows' by profiling and extending the Action class of ODRL.

This recommendation is the same as the recommendation from the previous Scenario (#3).

Specify ODRL Assets

This Scenario underlines the importance of finer grain information on ODRL Assets. An ODRL policy and rule constraining the access of a distributed analytic workflow may in some cases want to constrain the software Agent to accessing parts of a dataset. For example the policies and rules may want to constrain a software Agent to only query a limited set of variables within a data Asset. A data Asset can be identified in ODRL by a URI, so it is possible to get started using ODRL functionality as it is, but that arrangement requires each Policy

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

maker to specify sub-sets or sub-structures of their data Assets out of the model and therefore outside the interoperability sphere.

It is recommended that ODRL be used for these use cases, but refinements of Assets (using CDI variable model) be developed for distributed analytics (as part of a CDIF profile).

In a federated learning approach to data reuse, ODRL permissions and prohibitions need to consider the structure of the approach to model building and privacy preservation as part of the policy action.

6.4.3.5. Connecting CDIF and ODRL

CDIF recommends that a DCAT record be created, describing a dcat:distribution or, in lay parlance, a digital object, corresponding to the data set or resources to be used. For queryable data sources, this would instead be a dcat:service. (The dcat:dataset class is not used, because different distributions of it may have different access conditions.)

An ODRL Rule has a mandatory 'target' attribute which specifies the ODRL Asset to which the Rule applies. For example:

```
"permission": [
   {
      "target": "http://repository7987.org/dataset_6123",
      "action": "http://cdif.org/odrl/1/download",
      "assignee": "http://orciduser675765",
   }
```

Conversely, a DCAT record could express a predicate http://www.w3.org/ns/odrl/2/hasPolicy with the URI of the ODRL Policy as a target.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

This can be visualised as:

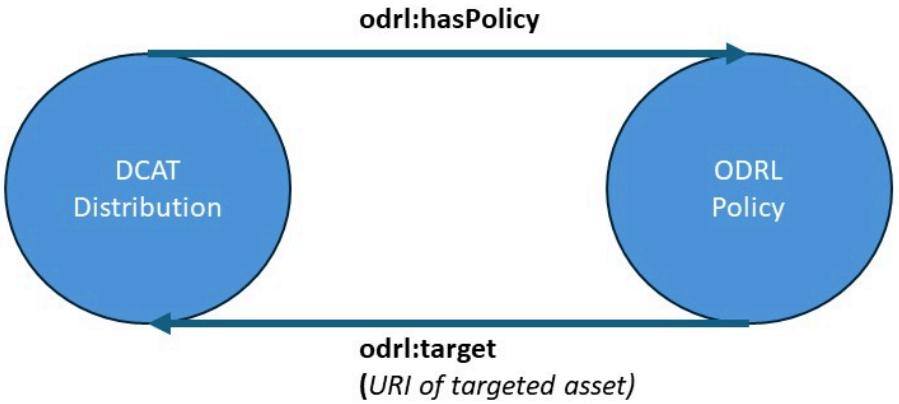

Figure 10. DCAT and ODRL - bi-directional.

There is a practical issue with implementing the above. An ODRL policy will typically have a one-to-many relationship with an ODRL Asset. That is to say, a repository may want to re-use one single ODRL policy for hundreds of datasets. One possible solution is to leverage the ODRL AssetCollection class, which can group multiple Assets under one URI. This in turn introduces other problems with defining persistent grouping rules that define an AssetCollection, such that the membership of the AssetCollection would not change unpredictably over time.

In the context of CDIF usage, a proposed but imperfect solution, is to populate the odrl:target attribute with what is essentially a global URI, the net effect of which is to make the above relationship unidirectional.

For example:

```
"permission": [
   {
      "target": "http://www.w3.org/ns/odrl/2/Asset",
      "action": "http://cdif.org/odrl/1/download",
      "assignee": "http://orciduser675765",
   }
```

leads to:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

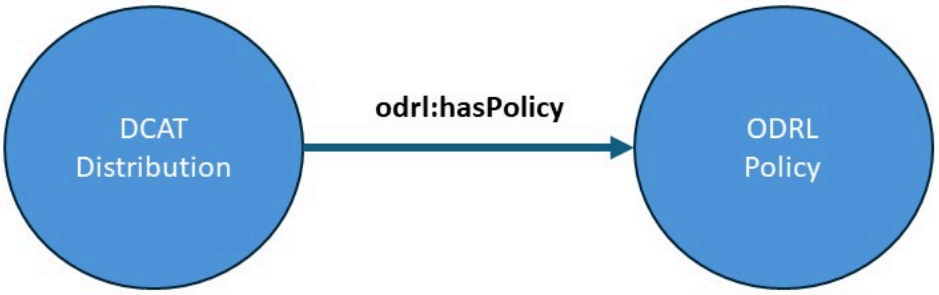

Figure 11. DCAT and ODRL - uni-directional.

6.4.4. Resources for this profile

ODRL Information Model: https://www.w3.org/TR/odrl-model/
ODRL Core Vocabulary: https://www.w3.org/TR/odrl-vocab/
Example ODRL access policies:
https://fiware-true-connector.readthedocs.io/en/latest/usage_control_rules.html
Arxiv preprint on applying ODRL in Transport Mobility Data Space: https://arxiv.org/pdf/2309.11289.pdf
ODRL Policy validator: https://odrlapi.appspot.com/
DUO to ODRL linkages:Pandit, Harshvardhan J., and Beatriz Esteves. "Enhancing Data Use Ontology (DUO) for Health-Data Sharing by Extending it with ODRL and DPV." *Preprint on webpage at https://www. semantic-web-journal. net/system/files/swj3127. pdf* (2023).
GA4GH Passport:

- ttps://www.ga4gh.org/news_item/ga4gh-passports-and-the-authorization-and-authentication-infrastructure/
- https://www.ga4gh.org/product/gdpr-international-health-data-sharing-forum/

GDI:https://zenodo.org/record/8208439
EJP-RD:

- https://www.ejprarediseases.org
- https://www.ejprarediseases.org/wp-content/uploads/2023/05/EJPRD_P2_AD49_PU_Virtual-Platform-Specification-V2.0.pdf

Comparison of ontologies for data asset-related permissions under GDPR:

- Esteves, Beatriz, and Víctor Rodríguez-Doncel. "Analysis of ontologies and policy languages to represent information flows in GDPR." *Semantic Web* Preprint (2022): 1-35 https://www.researchgate.net/publication/361160523_Analysis_of_ontologies_and_policy_languages_to_represent_information_flows_in_GDPR
- Semantics for implementing data reuse and altruism under EU's Data Governance Act. In: 19th International Conference on Semantic Systems, 20-22 Sep 2023, Leipzig, Germany. http://dx.doi.org/10.3233/SSW230015

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.5. Controlled vocabularies

6.5.1. Overview

Controlled vocabularies and related terminology-based semantic resources are highly significant from a FAIR perspective. They perform the essential role of defining the meaning of properties, types, values, or any other element in a volume of data, making them a critical component in scenarios involving (but not limited to) data integration and harmonisation. For the purposes of this document, we use the term 'controlled vocabularies' to cover all of the related set of terminology-based semantic systems, even though this may not be technically exact (ontologies, for instance, are often seen as a different kind of resource). Here, our use of the term potentially includes codelists, classifications, thesaurus, taxonomies, glossaries, ontologies, etc. In places where we specifically mean ontologies, classifications, or other types, these terms are explicitly used.

Some expressions of controlled vocabularies are richer than others in how they describe these systems of terms, concepts, and their relationships. A formal ontology or a statistical classification have more richness than a flat codelist, for example. For the purposes of CDIF we focus on the simplest expression of controlled vocabularies which will enable FAIR exchange, recognising that there may be more information available regarding any particular concept system than what is expressed according to the CDIF recommendations presented here. We limit our 'required' use of described controlled vocabularies to defining the enumerated values of fields within data sets and the codes used to notate them. This is not a comprehensive use of such resources, nor is it intended to be, but for the purposes of data description it is the only use on which other CDIF profiles have dependencies.

We envision two primary scenarios for FAIR usage of controlled vocabularies. The first scenario is a data-centric one: an application encounters a referrable term, code (that is, an encoding as opposed to software code), or symbol in a data set, and retrieves the necessary metadata to understand the meaning of that code through pointers provided by the description of the data structure. In this scenario, the navigation of the application is from the individual code or term to its meaning within the system of which it is part.

The second scenario is one where a controlled vocabulary is published as a reusable resource or consulted outside the context of a particular field in a data set, and is taken as a whole. In this scenario, navigation is from the controlled vocabulary down into its component parts.

Both scenarios require the atomic description of the controlled vocabulary, including the ability to navigate between the controlled vocabulary as a managed whole and its component parts. This has implications for how controlled vocabularies are represented, and how identifiers are used: each concept (node or term) must have its own identifier. This is especially critical in cases where data sets are to be integrated: a machine must be able to detect when the same concept is being used in two different data sets.

We recognize that there is an on-going discussion around the best PIDs to use for different purposes, and about how URIs should be formulated. In different domain and national contexts, there may even be regulations or standards which must be used. For the purposes of promoting cross-domain use of resources,

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

however, CDIF intentionally remains neutral in these discussions. PIDs are required, but how they are formulated is something which will vary widely across user communities: CDIF can only recommend how they should be included in a description of controlled vocabularies.

Specifically, it is important that there be a unique mapping from a URI to a concept, and that the URI be resolvable on the Web. If this is done, then the concept can be reused in a way which eases the burden of data harmonisation: if two data sets use the same concept, by referencing the same URI, then there is no ambiguity. To achieve this, CDIF recommends the use of SKOS for serialising concepts, providing definitions and dereferenceable URIs.

CDIF does not address many aspects of how controlled vocabularies should be handled for the purposes of FAIR, but in general is guided by Cox et al. (2021) 'Ten Simple Rules for making a Vocabulary FAIR'[117]. CDIF recommends the use of SKOS (as described in this section) to align with Rule 6 regarding machine-readable formats for CVs.

6.5.2. The Simple Knowledge Organization System (SKOS) and the Web Ontology Language (OWL)

When an organisation wishes to create a machine-actionable form of the terminology that it relies on, two questions that may arise include: 1) how can we digitise this such that it brings the most value to both human and machine agents? and 2) How can we share this widely to promote reuse and semantic harmonisation?

The first instinct may be to aim as high as possible on the semantic ladder, and to create a formal ontology (typically expressed in OWL). However, few organisations are prepared to create formal logical artefacts. (For an example of what is required for the definition of useful formal ontologies, see the material at OBO Foundry[118]). In the majority of cases, the way an organisation defines their terms will not be 'formal' in the sense of the logical formality that ontologies require. That is, the relations (object properties) between terms will not have domain and range constraints created to support reasoning and other checks for consistency. In fact, it often is not advisable for an organisation to create such formal artefacts, as much essential nuance may be lost during the process.

In most cases, a better solution would be to digitise and share the terminology using the conventions of SKOS. SKOS focuses on thesaurus-level relationships between terms (e.g. whether a term, or 'concept' in SKOS, has a broader, narrower, or simply related meaning relative to another), and has few constraints on definitional forms. Each term has an identifier, and SKOS artefacts can be shared using linked open data norms. The FAO [AGROVOC][119] provides an example of a complex, multi-lingual, and multi-nationally managed SKOS-based resource, managed through the VocBench software. SKOS artefacts are typically more intuitive to human users and can offer concept collections to further ease human browsing and discovery.

[117] Cox SJD, Gonzalez-Beltran AN, Magagna B, Marinescu M-C (2021) Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17(6): e1009041. https://doi.org/10.1371/journal.pcbi.1009041
[118] https://obofoundry.org/
[119] https://www.fao.org/agrovoc/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

When such a SKOS-based resource is in place, the 'official', human-oriented definitional space is set and has a persistent presence on the Web (assuming the artefact has been shared through solutions aligned with the LOD and FAIR Principles). At this stage, variants of the official terms that are more logically formalised (and thus more machine-friendly) can be created in one or more ontologies, using the expressive capabilities of OWL. For example, the term [landslide] (https://www.eionet.europa.eu/gemet/en/concept/4668) in the General Multilingual Environmental Thesaurus (GEMET) has been formalised in OWL using the Open Biological and Biomedical Ontologies (OBO) Library guidelines in the Environment Ontology (ENVO) class for [landslide process] (http://purl.obolibrary.org/obo/ENVO_01000692).

CDIF recommends using SKOS as a solution to the question of how to provide a useful, machine-actionable representation of a controlled vocabulary. It should be emphasised that for data integration any given concept should be identified using its URI as assigned by the authoritative source, so that reuse of concepts can be known.

6.5.3. Using SKOS

In practice, this means that a dataset schema requires categorical values for a variable in the dataset to be members of a skos:ConceptScheme. Instance values for the variable would have to uniquely identify a concept from the skos:ConceptScheme that specifies the set of possible variable values.

To do this, a limited number of SKOS classes and properties are used, as described below. As the lowest common denominator, we will refer to any enumeration of values as a 'codelist' here, even if it is a richer structure when described according to some other model (such as a classification in the DDI-CDI model).

To use SKOS in an interoperable fashion in CDIF, we must have conventions about how the concepts it describes are referenced. Because we use SKOS concepts in a very limited fashion in CDIF, this convention can be expressed as a fairly simple (albeit technical) rule:

> In the CDIF Data Description profile, which uses the DDI-CDI Model, connections are made in RDF instances using the 'takesSubstantiveValueFrom' and 'takesSentinelValueFrom' associations of the RepresentedVariable class, relating it to one or more skos:ConceptSchemes. This indicates that any skos:Concept in that scheme provides the equivalent of a cdi:Category, with its skos:Notation property containing the value of a cdi:Code's associated cdi:Notation TypedString property, as contained in the string value associated with that Notation's DDI-CDI complex data type.

This has the consequence that SKOS ConceptSchemes used for the purpose of describing enumerations appearing in the data *must* have a skos:Notation, and that they may not have more than one. The exception here would be those cases where the concepts URI is used in place of some other notation, in which case the concept URI can be used directly in place of the value of the SKOS Notation.

ConceptSchemes: In this approach, a skos:ConceptScheme is a set of concepts that define the possible values for a categorical variable. The scheme can define a hierarchy of values from general to more specific. This hierarchy is represented using the skos:broader relationship linking more specific concepts to subsuming

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

more general concepts. If a hierarchy is defined in the scheme the top concept(s), i.e. those that do not have any skos:broader associations, must be identified using the skos:hasTopConcept property on the skos:ConceptScheme.

Concepts: SKOS Concepts are used to represent the possible values for a categorical variable. In the RDF implementation of a skos:ConceptScheme, these are the requirements for each concept:

- must specify rdf:type skos:Concept.
- must have a unique identifier. Ideally this is a globally unique, resolvable URI or PID.
- must have a skos:inScheme relationship to the containing skos:ConceptScheme.
- must have at least one skos:prefLabel, but may have multiples as these are language specific (SKOS does not permit more than one skos:prefLabel per language-locale).
- must provide a definition of the concept with skos:definition, also in language-specific form.
- If the concept scheme is hierarchical, skos:Concept must use the skos:broader relationship to indicate its parent in the hierarchy,

Various other properties can be provided

- Other labels may be provided using skos:altLabel, also in language-specific form.
- skos:narrower relationships (inverse of skos:broader) can be provided, supporting navigation both up and down the hierarchy.
- A unique (in the scope of the vocabulary) skos:Notation must be provided to denote the skos:Concept in data instances. Such notation values are commonly short strings or abbreviations that are easier for users to interpret than the concept identifier. (Note the exception above, where unique URIs may be used instead of a skos:Notation).

This use of SKOS materially aligns with that described in the document 'Modelling of Eurostat's Statistical Classifications in ShowVoc' for classification items.[120]

6.5.4. Formal statistical classifications

Formal statistical classifications have some additional information which may be useful to describe. The recommendation is that this be done according to the style used at Eurostat[121], FAO[122], and increasingly by other organisations in the UN family.

Note that such descriptions include a richer set of properties, and may involve a description of mapping tables between versions of classifications. Such relationships are modelled using the XKOS specification[123], for which there is a user guide.[124]

[120] https://cros.ec.europa.eu/book-page/modeling-eurostats-statistical-classifications-showvoc
[121] https://cros.ec.europa.eu/book-page/modeling-eurostats-statistical-classifications-showvoc
[122] https://www.fao.org/statistics/caliper/resources/data-modeling/en
[123] https://rdf-vocabulary.ddialliance.org/xkos.html
[124] https://linked-statistics.github.io/xkos/xkos-best-practices.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.5.5. Mappings between controlled vocabularies

For general mappings between controlled vocabularies, there is increasing use of 'A Simple Standard for Sharing Ontology Mappings' (SSSOM)[125]. This style of mapping is somewhat limited, as it does not necessarily provide all of the information needed to automate a transformation between the values of two variables encoded with different-but-corresponding values. It is a good starting point, however.

As of this writing, a group is being formed in RDA to specifically address issues around mappings, the 'FAIR Mappings Working Group', which came out of a discussion at International Data Week in Salzburg in 2023.[126] This working group will be the place where good practice in this area is likely to emerge, as it involves participants responsible for many related efforts within FAIR Impact[127], EOSC (the Metadata Schema and Crosswalk Registry[128]), and in other initiatives.

6.6. Data integration

This section lays out the process of describing data to be disseminated in an 'integration-ready' form. The general intention is that domain standards containing much of the needed metadata can be mapped into the 'integration-ready' form for exchange across domain and infrastructure boundaries, and that this mapping can be done with an acceptable degree of effort, in some cases automated completely. Some examples of how these transformations can be achieved are provided. A further section explains how the production of integrated data sets can be described, although CDIF does not provide recommendations at this point for doing so.

This discussion is focused on data that can be represented in tabular formats with simple literal values. While this does not cover all the data we wish to describe, it represents a significant portion of it. Much of the data made available for reuse is expressed in CSV or similar, text-based formats, and these provide the initial focus for describing data. In future, this coverage will be expanded: this initial focus is intended as a useful starting point.

6.6.1. Introduction

Data integration is a task performed by researchers and others who wish to produce a unified data set which can be subject to analysis, or more easily used in analysis. Many of the decisions made during the integration of two unlike data sets depend on the analysis to be performed. For CDIF, the main goal is to provide a description of data so that it is 'integration ready'. In a FAIR scenario, the user of the data is not a known quantity, and only they can make the decisions about how to integrate the data they are working with.

There is, however, a secondary scenario in which a data integration has been performed for presentation of harmonised data to a user. In this case, the fashion in which the data of interest has been integrated from various sources is a critical aspect and needs to be described, and the process steps recorded - this crosses

[125] https://mapping-commons.github.io/sssom/
[126] https://www.rd-alliance.org/lets-talk-about-fair-mappings-towards-common-practices-sharing-mappings-and-crosswalks/
[127] https://fair-impact.eu/events/fair-impact-events/documenting-mapping-community-practices
[128] https://faircore4eosc.eu/eosc-core-components/metadata-schema-and-crosswalk-registry-mscr

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

into what we often think of as ‘provenance’ metadata, which covers a detailed description of the processing and transformations performed, along with information about the methods, etc. In this initial description, we do not attempt to provide this more-complete description.

Both of these scenarios require a wealth of metadata at a granular level if we are to maximise the automation of data integration functions. This is an important goal, as the cost of ‘data wrangling’ is high in terms of overall resources. While it is likely that data integration will always require some attention and input from the researcher, many of the necessary tasks are routine and can be automated if sufficient information is known about the data themselves. CDIF attempts to provide a sufficient level of metadata to support such automation.

Given this granular description of the data, it is possible to extend this information set to describe the mappings used to merge data sets. This information can be expressed at the level of the enumerations and types of data found in individual fields of the data sources, both in terms of their meaning (semantics) and in terms of their syntactic representation. In addition, other processes may be in play during a data integration. CDIF separates these different aspects, providing for a description of that individual data sets, in reference to the controlled vocabularies which provide the semantics and codes used in the data, and supplementing these descriptions with separate expressions of the mapping between the controlled vocabularies used, and with a more-complete description of the processes which were performed. Together, these aspects make up the provenance of the integrated data set.

The CDIF recommendations for data integration focus on two specific areas:

1. The description of data sufficient to make it ‘integration ready’, including references to the needed controlled vocabularies for describing field-level semantics and enumerated values;
2. The description of how two or more integration-ready data sets can be combined to describe an integration process, through the inclusion of mappings and process descriptions. (The recommendation for describing controlled vocabularies, mappings, and processes are treated separately: here we explore how these different metadata are used in combination to describe a data integration).

For the first scenario, we make specific recommendations about data description. The second area is one where more work is needed, but some progress has already been made. This description of data integration is a priority activity for future versions of the CDIF recommendations.

6.6.2. Problem statement

In a cross-domain scenario, it is necessary that a data resource, once located, can be loaded into the systems which will make it practically available for use. This is a metadata-intensive task, and one which has several dimensions. CDIF attempts to separate the metadata requirements in a fashion which allows them to be more easily satisfied.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The disseminator of data cannot know what structure will be needed by potential users, as their integration requirements and systems may be designed to work with data which is structured differently. The disseminator can, however, describe the data as they manage and present it, along with information about its logical contents. The user can then re-structure the data as needed for their own use, and do so programmatically. Sufficient metadata must be available to support this programmatic restructuring, without losing any of the information about the data - especially its links to semantic definitions.

The process for providing such detailed descriptions of the data can be broken down into a series of steps:

1. Identify the logical variables in the data, where each 'variable' measures a single characteristic of a single unit type, using a consistent set of values. The possible values must be enumerated or otherwise described in a detailed fashion. Representations must be able to identify domain-agnostic semantic descriptions for each possible value, and the variable definitions themselves must similarly be independent of any domain specificity.
2. Indicate how the logical variables fit into the structure of the file, by specifying also any 'presentational' variables (that is, variables which may be needed for structuring the file, rather than for describing the logical content of the data). Relationship between presentation and logical variables must be specified. The complete set of variables can then be described as a 'logical record'.
3. Describe which fields are used to identify the records (the 'primary key').
4. Describe the encoding of all variables physically present in the file, and how they are sequenced and stored for programmatic retrieval.

The description of variables includes an indication of where the semantics for both the field and the values come from. Because the semantic distinctions used in different domains can be very diverse, it is often not the case that these can be automatically subject to mappings or transformations. This step often requires the judgement of the user. Presenting the semantics used in the data — programmatically mapped/transformed to the extent possible — is often the closest we can come to a fully automated data integration.

Even so, being able to automate the structural transformation of data, accompanied by the needed semantics, with the roles they play vis-a-vis the data clearly defined, represents a major efficiency gain: manually performing these steps can be costly and time consuming.

6.6.3. Standards landscape

Identifying standards which will meet the requirements described above is a challenge. Many domain-specific standards have excellent descriptions of the data used within their domains, and some of these separate the semantic aspects of data description from the structural ones in the fashion described above. The number of candidates is small, however.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

CSV on the Web[129] is one candidate, as is the W3C's *Model for Tabular Data and Metadata on the Web*[130]. *Frictionless Data*[131] provides a similar lightweight description of data. More metadata-rich models also exist, such as the SDMX[132] Information Model and the RDF DataCube Vocabulary[133] which is based on it, but these are largely limited to describing multi-dimensional data. Although the *DDI Codebook*[134] and *DDI Lifecycle*[135] specifications are very close to the CDIF requirements, they are also inherently bound to their intended use for describing social, behavioural, and economic (SBE) data. Only the related *DDI Cross-Domain Integration* (DDI-CDI) specification[136] meets all of these requirements, largely because it was designed specifically to describe cross-domain data for the purposes of integration.

It is worth considering the strengths and weaknesses of each of these potential models, so that we can better understand the problem and the metadata needed to solve it.

We must have sufficient agreement between systems so that a requesting system can obtain and transform the data described, based on the disseminator's description of the data as understood within the distributing system, and to do so without benefit of human intermediation. Thus, we cannot safely rely on the conventionalized use of very flexible formats. On the other hand, we do not want to be overly restrictive in how data is described, as this will increase the barriers to use by disseminators.

This is a difficult challenge indeed, and it accounts for the fact that this type of ubiquitous, generic data description has been slow to emerge. To support cross-domain FAIR, as envisioned by CDIF, however, we must have a workable solution.

Consider the different specifications listed above. All of these provide a separation of concept definitions - semantics - from the structural description of data, a pre-condition for a useful cross-domain standard.

CSV on the Web is designed to add metadata to CSV tables, which is a use case not dissimilar to the one we address in CDIF. The problems arise when we consider the flexibility of the CSV format itself: there are no limitations in CSV regarding the logical organisation of the contents of the table: is each row a unit being observed, and each column a characteristic of that unit? Or are the columns instead the units? This logical certainty is required. Although CSV on the Web allows us to add richness to the meanings of columns and rows of a table, it is in this sense too flexible to make a good foundation for the scenario addressed by CDIF. We can know what concepts are in play, and how rows and columns are organised, but we cannot understand how the presentation of the data — the table organisation — is *logically* structured. What roles do the concepts play? While this could perhaps be reverse-engineered, it is easier to explicitly state how the concepts logically relate, and how these in turn are presented in a tabular form. A similar critique can be applied to *Frictionless Data*'s data description, and to the *Model for Tabular Data and Metadata on the Web*.

[129] https://csvw.org/
[130] https://www.w3.org/TR/tabular-data-model/
[131] https://framework.frictionlessdata.io/docs/guides/describing-data.html
[132] https://sdmx.org/
[133] https://www.w3.org/TR/vocab-data-cube/
[134] https://ddialliance.org/Specification/DDI-Codebook/2.5/
[135] https://ddialliance.org/Specification/DDI-Lifecycle/3.3/
[136] https://ddialliance.org/Specification/ddi-cdi

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

These specifications are all designed to perform more limited tasks, and they do it well (they were all inputs in the design of the DDI-CDI model, in fact), but they fail to cover all the needed logical relationships, being restricted to tabular descriptions which combine logical and presentational aspects of the data.

SDMX and the *RDF DataCube Vocabulary* have a different issue. They have stronger formal models for how concepts intersect with data structure, and do not combine the presentational and logical aspects of data description in the same way, but they are limited because they insist on a multi-dimensional description of data sets. (It should be noted that while SDMX 3.0 has introduced a 'microdata' description feature, this is still new, and is not yet the version adopted by most implementations, nor is it the version on which the *RDF DataCube Vocabulary* is based.)

SDMX demands that the metadata description be provided using a model which may or may not be supported by disseminator's systems. Many data repositories do not manipulate data as multi-dimensional cubes, and lack the information needed to describe their data in this fashion. Further, *SDMX* uses a very disciplined cube definition: all data has regular dimensionality. In the real world, even many systems which are based on multi-dimensional models have irregular dimensionality and 'sub-cubes', which are not permitted in *SDMX*. Given that *SDMX* is an exchange model for official statistics, this disciplined approach is very reasonable, but is not appropriate for CDIF, where existing data stores must be described according to metadata which already exists.

Both *DDI Codebook* and *DDI Lifecycle* are excellent models in terms of being generic-but-concept-rich data descriptions, but they also lack the range of support needed for CDIF. In the SBE realm, data is overwhelmingly stored and processed as 'wide' data files: unit record data, where each record is a set of measurements or values about a single unit, one per row. (Admittedly, this is a simplification, as there can be hierarchical relationships between such records, but that is a side point for the purposes of this discussion.)

The *DDI Cross-Domain Integration* specification was developed exactly because it is increasingly common for other types of data to be combined in research projects with this traditional form of SBE data. Four different types are identified: wide, long, multi-dimensional, and key-value (the kind of data commonly found in 'big data' systems). Because they are described in a single model, intended to support transformations between these different structural types, the needed separation of logical and presentational aspects of the data was a necessity.

One might ask why a graph model is not used instead, as graphs do a complete separation of presentation and logical relations, and are absolutely flexible. (Indeed, the virtues of graphs are exactly why CDIF uses them for the purposes of structuring metadata.) It is not common practice to store or disseminate data in graph form today, for several reasons. Graphs are difficult to manage, especially when historically data has been managed in 'data sets' stored as files or tables in databases. While some domains are able to manage and disseminate data at a very granular level, this is certainly not true of all. Further, this type of system places demands on the scalability of systems, and the RDF solutions available today for disseminating graphs often fail when it comes to performance. For metadata, which tends to be smaller in size, this is less of an issue, But if a URI must be assigned to every single value in each organisation's data repositories, the

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

management burden quickly becomes a barrier. While such an approach may become feasible in future, it is not feasible across domains at the current time.

In DDI-CDI terms, graphs of data are essentially key-value systems: each data point has a unique identifier, allowing it to take its place within the graph. The DDI-CDI model accommodates this approach conceptually, as well as the way it has been implemented in big data stores, but this is not sufficiently common to be used as the basis for a more limited description of data for the purposes of FAIR data exchange and integration as intended here.

Ultimately, while DDI-CDI may seem to be a complex model, attempts to produce a 'simpler' model based on existing W3C specifications produced equal complexity, and were sufficiently extreme to no longer be extensions of another standard, but effectively new specifications in and of themselves. Using the DDI-CDI specification seemed to be the right approach for CDIF, which is intended to recommend other specifications, not to produce what are practically new ones, even if based on existing, simpler specifications designed for other purposes.

6.6.4. Implementation examples

CDIF is intended to leverage metadata which exists already with the systems and standards used within domains and communities of use, allowing it to be expressed as CDIF descriptions to better support reuse by others outside those communities or domains. Data description at a granular level sufficient to support interoperability and reuse is a metadata-intensive activity - one of the biggest challenges of FAIR implementation - so it is important to understand in practical terms how much effort will be required to support it. This section describes some of the early examples where the DDI-CDI profile used in CDIF has been implemented. (Some further thinking in this direction in the context of EOSC can be found in the paper 'The Role of DDI-CDI in EOSC'.[137])

Three examples are described here to illustrate the kinds of approaches which are anticipated. The first is a demonstrator tool which was designed for an exploration into how the SDG Indicators and their disaggregations could be easily described for integration. While the SDG Indicators are disseminated in SDMX, there was no similar publication mechanism for the disaggregations, although this topic is one which receives a lot of attention.[138] The tool was designed to take a simple CSV expression of the data and allow it to easily be marked up through addition of the needed additional metadata.

Our second example is one which was developed as part of the work in the WorldFAIR Social Surveys Case Study (WP06). In that domain, Stata and SPSS are very common software packages for data processing and analysis. The tools developed to mine the data and metadata out of these files and express them in DDI-CDI.

[137] Gregory, A., Hodson, S., Wackerow, J. (2021). The Role of DDI-CDI in EOSC: Possible Uses and Applications. Zenodo. https://doi.org/10.5281/zenodo.4707263.
[138] See https://unstats.un.org/sdgs/iaeg-sdgs/disaggregation/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Our third example is a prototype developed to pull data and metadata from the SDMX Web services offered by ILOSTAT[139] and describe it in DDI-CDI format, with the idea that this could be used for further integration.

These examples show how existing resources can be used to generate those required to support the production of CDIF 'integration-ready' data with a reasonable amount of effort. Details on each example are provided below.

6.6.4.1. CSV-to-DDI-CDI Sample Generator example

This tool takes a CSV file (you can select one of the pre-loaded examples or upload your own) and generates the DDI-CDI metadata in the form of JSON-LD (as recommended by CDIF, although to an older version of the profile). Not all of the needed metadata is available in the CSV, and the tool provides an interface for a person to add the additional definitions and descriptions. It can be found at https://ddialliance.github.io/ddi-cdi-sample-generator/.

CSV is a format which is easily exported from many different systems, and is often used as an interchange format, In such cases however the structure and content is often required to be in some conventionalised form to support interoperability with other applications. This tool takes a very common style of CSV for aggregate data and allows it to be described with the needed additional metadata. The prototype tool runs as a stand-alone javascript application in a browser; there is no required server component. Thus, it is an extremely light-weight type of application.

The importance of this prototype is that it illustrates how an organisation which does not have sophisticated systems could manually produce CDIF-compliant data descriptions. While this would require some effort, it is not greater than what data archives and repositories do today when implementing standards such as DDI Codebook (the DDI Codebook expression of the metadata in XML is also produced by this tool, although not as useful for cross-domain FAIR purposes.) It does provide an easy way for developers of more sophisticated systems to generate sample code for testing and exploratory purposes. The tool here is not a production tool, but there is currently some discussion within the DDI Developer's Community of producing a production version.

6.6.4.2. SPSS/Stata-to-DDI-CDI Converter example

This is another exploratory prototype, developed by staff at Sikt, who are active both within WorldFAIR and in the DDI-CDI Working Group. This prototype will be a deliverable from the Social Sciences Case Study of WorldFAIR. It produces the XML expression of DDI-CDI, but this could as easily be rendered in JSON-LD, as per the CDIF recommendations (the two syntax representations are based on the same model, being generated programmatically from it.) It can be found at https://ddi-cdi-converter-app.azurewebsites.net/.

This tool also provides a way for a person to manually add metadata which is missing from the SPSS or Stata files. Unlike the CSV converter, this tool is server-based, and supports the description of Wide Data Sets, rather than Dimensional Data Sets. (The two are quite similar, however).

[139] https://ilostat.ilo.org/resources/sdmx-tools/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

This is also a prototype, but it illustrates the way in which data archives in such domains as Social Science and Public Health — which often have data in SPSS and Stata formats — could produce 'integration-ready' descriptions of their data holdings to support cross-domain FAIR use, in line with the CDIF recommendations.

6.6.4.3. ILOSTAT SDMX-to-DDI-CDI example

The Statistical Data and Metadata Exchange (SDMX) standard is used widely within the official statistics community for the publication of aggregate dimensional data (and related metadata) on a wide range of topics. It is concept-rich, and is accompanied by a body of harmonised concepts which support data exchange. Moreover, SDMX provides standard application programming interfaces (APIs) as RESTful services.

Within the international statistical system there are a very large number of important aggregate dimensional data sets that are readily available in SDMX. The SDG Indicators are a primary example. Similarly, the International Labour Organization (ILO) publishes statistics on labour using SDMX, covering a wide range of data used for policy at the national and supra-national levels, as well as forming a large body of important reference data for scientific research.

While SDMX is excellent for structured data exchange,and includes metadata standards that are robust within the context of aggregate statistical data, it may not fully address the needs for integrating data across different domains and formats, and its metadata scope may be limited when dealing with more detailed or domain-specific information. Moreover, SDMX's standardised structure, while beneficial for consistency, can sometimes be rigid when attempting to extend beyond its predefined concepts and dimensions.

On the other hand, DDI-CDI is designed for cross-domain data integration, facilitating the combination of data from various sources and formats. It supports detailed data documentation and the use of standardised metadata, which enhances data interoperability and usability in broader contexts, including scientific research. In addition, DDI-CDI can describe data at a more granular level, including microdata, longitudinal studies, and complex datasets. It is designed to be more flexible and extensible, allowing for the inclusion of additional metadata elements and the adaptation to various data types and structures, making it suitable for integrating statistical with non-statistical data.

This section presents as an example a prototype developed at the United Nations Statistics Division of the UN Department of Economic and Social Affairs, exemplifying how the standard SDMX APIs could be mined to describe dimensional statistical datasets using for statistical data and their related metadata, which could then be transformed them into a minimum DDI-CDI profile, with the objective of facilitating their JSON-LD for further use in further data integrations.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

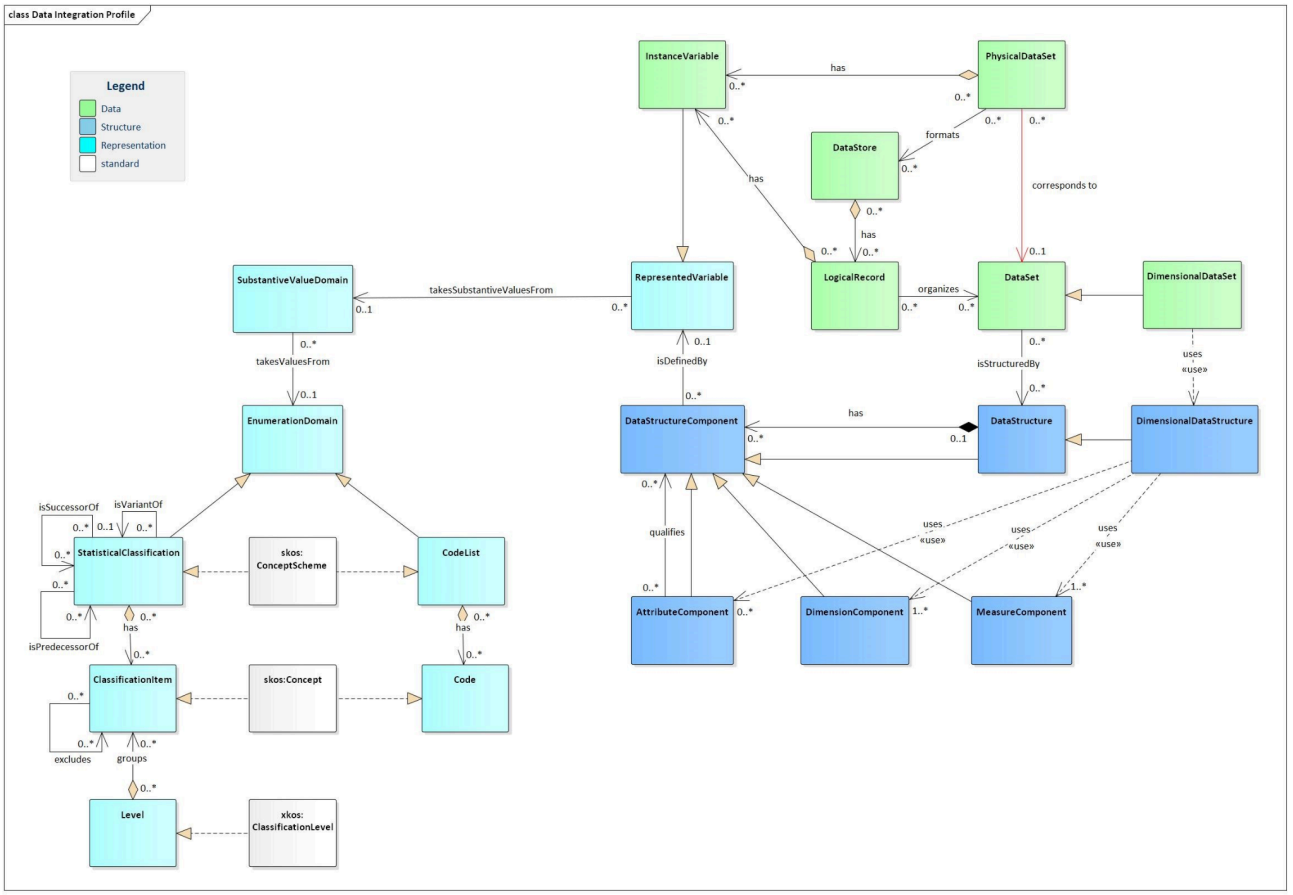

Figure 12. Prototype of a minimum DDI-CDI profile for data integration purposes.

This example is more fully documented than the other examples, with a tutorial available alongside the code in a Jupyter Notebook:

- Github link: https://github.com/Cross-Domain-Interoperability-Framework/undata_example_cdif
- ILO Jupyter Notebook link: https://github.com/Cross-Domain-Interoperability-Framework/undata_example_cdif/blob/main/src/cdip1_ILO_example-v2.ipynb

SDMX is used widely within the official statistics community for the publication of aggregate data on a wide range of topics. For this data, SDMX is the standard used both for data and metadata. Primarily designed to describe aggregate statistics, SDMX is concept-rich, and is accompanied by a body of harmonised concepts which support data exchange. SDMX provides standard application programming interfaces (APIs) as RESTful services.

The SDG Indicators are a primary example of this type of data, but within the international statistical system there are a very large number of important data sets which are of a similar type. The International Labor Organization (ILO) is responsible for the publication of statistics on labour within a system that covers a wide

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

range of data used for policy at the national and supra-national levels, as well as forming a large body of important reference data for scientific research.

This prototype takes data from the ILO implementation of SDMX Web services as an example, but this approach could be easily applied to a large number of other SDMX-using organisations within the international statistical community (Eurostat and the European Central Bank are also major users of SDMX, for example.)

The implications for a cross-domain FAIR scenario here are large: the body of official data, used as a reference point in many scientific domains, can be easily made available for integration with other research data. The SDG Indicators - one of the many data sets available through SDMX APIs - would become available for use not only on their own, but in combination with any other data as the basis of integrated data sets for analysis.

6.6.5. Describing data to make it 'integration-ready'

6.6.5.1. CDIF data description steps

This section explains how to describe a data set or a service providing data so that it can be used for integration and similar purposes. CDIF uses a subset of the classes in the DDI-CDI specification to do this. We will introduce the classes needed for each step, with more detail provided on the full set of needed properties in Appendix 4.

For a data set — that is, a static set of data — there are four steps. For a service, where the structure and physical format of the data will depend on the service, the last two steps are not required.

The steps are:

1. Describe the Data Set or Service;
2. Describe the Variables;
3. Describe the Data Structure;
4. Describe the Physical Format of the Data.

6.6.5.1.1. Describe the data set or service

The class used here will depend on what is being described. For Data Sets, the cdi:DataSet class is used. For services — that is, queryable sources of data — the cdi:DataStore class is used. The difference between them is that a Data Set object in DDI-CDI is static, with a fixed set of records (similar to what is seen in DCAT for a Distribution), a Data Store is more flexible: while it always contains the same logical records, their expression (and the set of units measured) may change. These can be seen in the diagram below:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

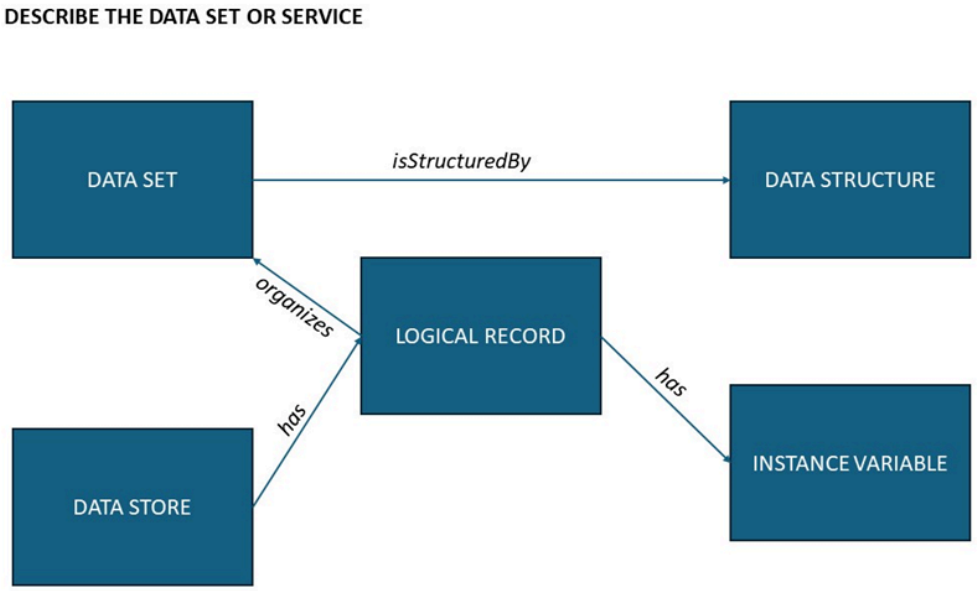

Figure 13. Describe data set or service.

A Data Set and a Data Store both have Logical Records: arrays of variables which make up the records in the data. The Logical Record has a set of Instance Variables, described more completely in the next step. These are the different fields which are used to hold the data (like the columns in a SQL table). A Logical Record is not a physical arrangement of the data – it is simply a set of related fields, all of which will potentially exist in a single record.

An additional class may be associated with the Logical Record if desired: a Population, Universe, or Unit Type may have an *isDefinedBy* association with it. These classes are concepts which can be used to associate the Logical Record with the thing it measures. The use of this class is always optional: we will describe these when we talk about variables, below.

A Data Set has a Structure, which can be described — a Data Store (a ‘service’) does not have such a fixed structure, as this will generally be dependent on how it is queried. The Structure description varies according to what type of data set is being described — whether it is organised as a 'wide' table, as a ‘long’ stream of data, or as a cube. Each of these types of data structure will be described separately.

While not shown in this diagram, the variables will be used as ‘components’ in the description of the data structure. A component is the role played by a variable in the context of a specific data structure: it is an identifier? A measure? A descriptor? The use of components for different kinds of data structures is detailed for each type of data structure in turn, as these are discussed below.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.6.5.1.2. Describe the variables

In order for data to be easily reused, it is important for both users and machines to have details about the variables themselves. In DDI-CDI, there are several types of variables, each performing a different function. (This is termed the 'variable cascade', and may be of interest, but is out of scope for our immediate purposes.) Only two of the DDI-CDI variables are used here: the cdi:InstanceVariable class and the cdi:RepresentedVariable class.

The Instance Variable is the variable as it is used in the context of a Data Set. The Represented Variable is a variable used as a template across multiple Data Sets or Data Stores. It is worth considering the value of providing this additional metadata around Represented Variables.

To describe a single data set, we do not need to provide information on Represented Variables - the fact that a variable is reused elsewhere is not necessary for someone to process the single data set. However, in cases where a data set re-uses a variable to promote comparison between data sets, this becomes very important. In many scenarios, the same variable is used in many different data sets to provide a point of comparison or integration. A typical case is that of repeat data collection across time, where we want to take a series of measurements to produce a time series.

Often, although the measurements are identical, they are given different names for the purposes of data management, typically by appending the time of data collection (such as a year) or a number indicating the wave (e.g., a repeated measurement of education level might have variables 'educ_level_95', 'educ_level_96', 'educ_level_97', and so on). In such cases, a Represented Variable can be extremely useful because it indicates without any ambiguity that two variables are in fact the same. This information can be very valuable to users performing integration or harmonisation of the data, or producing their own times series from it.

The DDI-CDI model uses Represented Variables when describing structures, and for many other purposes, but uses Instance Variables when describing physical layout of the data records. This can seem confusing. In the DDI-CDI model, we can avoid the complexity of this arrangement, however, because according to the DDI-CDI model the Instance Variable is a sub-class of Represented Variable, meaning that any given Instance Variable can also be treated as a Represented Variable, and inherits all its associations and properties.

For our profile, Represented Variables are always optional. When present, they should be declared *in addition* to Instance Variables, and the Instance variables which are based on them should have a *uses* association declared with the corresponding Represented Variable. Their representations will always be the same. In the diagrams, Represented Variable is presented in grey, to indicate this special, optional status.

In the diagrams below, we always show the Instance Variable being used, as that is required. Note that if Represented Variables are present, these may be the ones which appear in the full DDI-CDI model, as for describing the representations of variables and their use as components in structures.

The diagram below shows the basic class structure for variables:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

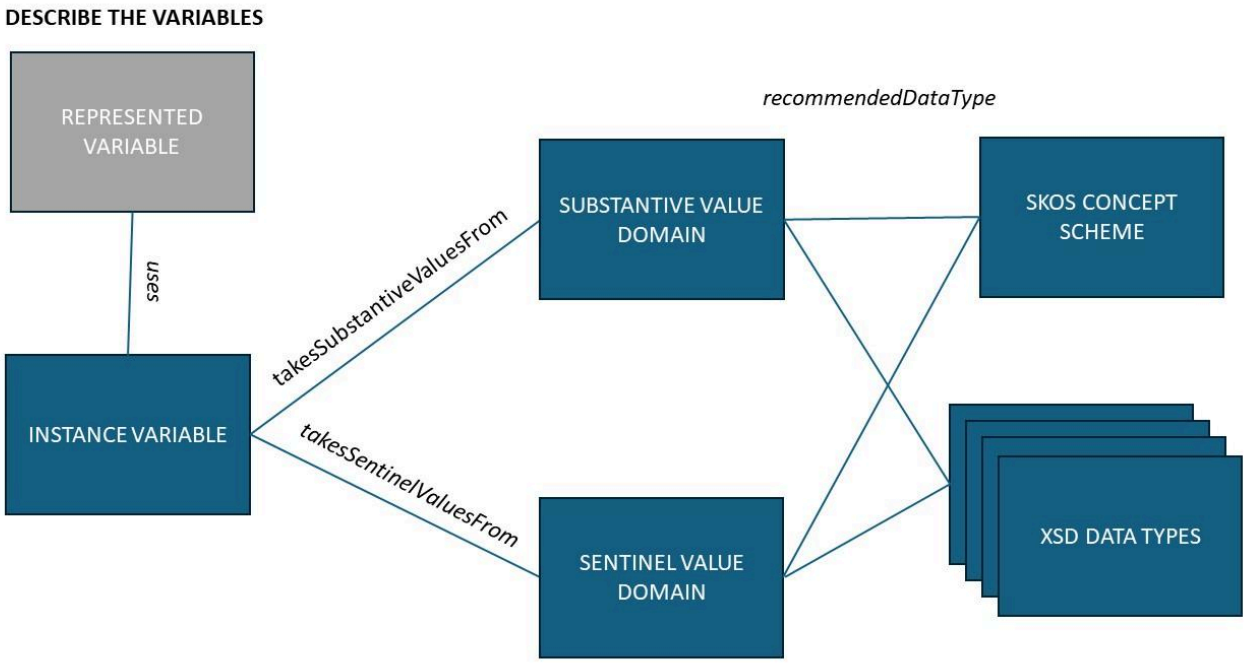

Figure 14. Basic class structure for variables.

Any given variable has a set of possible values, each of which comes from either a Substantive Value Domain (the meaningful values) or a Sentinel Value Domain (the flags for things like “missing” values). It is required to have an instance of the cdi:SubstantiveValueDomain class when describing a variable. It is not required to have an instance of the cdi:SentinelValueDomain class, although it is recommended if these are present in the data, as this enhances the practical reusability of the data. It is always the case that all possible values for the variable be described in one or the other of the value domains, however.

Both the Substantive Value Domain and the Sentinel Value Domain take their data types from either a reference to a skos:ConceptScheme (for enumerated values) or an XSD data type. The list below shows the valid XSD data types (XSD types available in RDF – see https://www.w3.org/2011/rdf-wg/wiki/XSD_Datatypes):

xsd:anyURI
xsd:base64Binary
xsd:boolean
xsd:byte
xsd:date
xsd:dateTime
xsd:decimal
xsd:double

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

xsd:float
xsd:gDay
xsd:gMonth
xsd:gMonthDay
xsd:gYear
xsd:gYearMonth
xsd:hexBinary
xsd:int
xsd:integer
xsd:language
xsd:long
xsd:Name
xsd:NCName
xsd:NMTOKEN
xsd:negativeInteger
xsd:nonNegativeInteger
xsd:nonPositiveInteger
xsd:normalizedString
xsd:positiveInteger
xsd:short
xsd:string
xsd:time
xsd:token
xsd:unsignedByte
xsd:unsignedInt
xsd:unsignedLong
xsd:unsignedShort

The definitions of variables are described in Appendix 4, below — these are provided as literal definitions but can also make reference to concepts defined elsewhere. Note that some variables may hold more complex values such as arrays. We do not cover this case in the current profile, but this is a subject which will be addressed in future.

It is also possible to instantiate classes which represent the subject of the data. This is possible at both the variable level and the Logical Record level. These classes include cdi:UnitType, cdi:Universe, and cdi:Population. A Unit Type is the most general class: it differentiates the top-level category of entity being observed, and is often implicit in the more-detailed classes (Universe and Population). Typical Unit Types are ‘person’, ‘organism’, ‘chemical compound’, ‘water’, and similar very broad categories. While conceptually always present, it is not always useful to describe these, as they are often implicit in the more-detailed (and more useful) Universe and Population descriptions. A Universe is a qualified Unit Type, but without specifics about time and place. For a Unit Type of ‘person’, I might have ‘student’ as a Universe. For a Unit Type of ‘organism’ I might have ‘dipterids’ as a Universe. A Population further qualifies the Universe by specifying

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

time and place. These classes are instantiated with formal, textual definitions, which may be included as literal or by reference to an external source.

The diagram below shows how the variables and Logical Records can be associated with Unit Types, Universes, and Populations. The use of Unit Type, Universe, and Population is always optional, but may be useful, especially when variable-level searches are being made across broad sets of data for comparable measures on comparable populations.

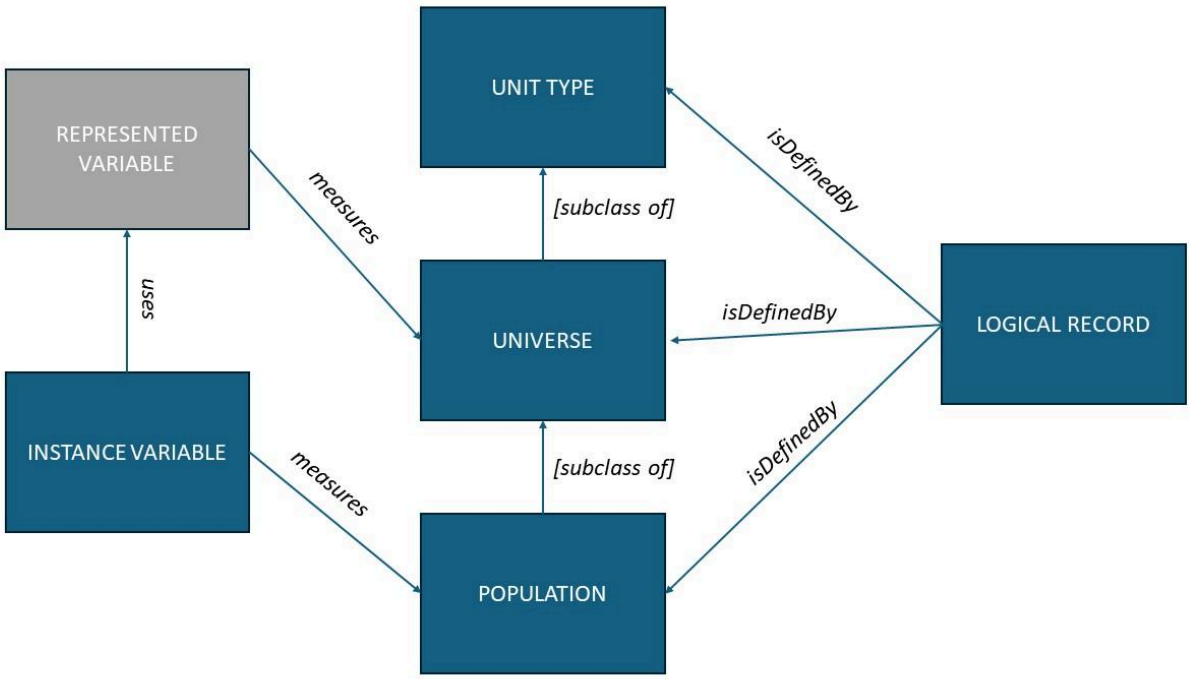

Figure 15. Unit, universe, and population.

6.6.5.1.3. Describe the data structure

For Data Sets (which always have a fixed structure – and these may include the results of querying services, with the Data Structure generated when the query is executed along with the resulting data), DDI-CDI gives us a model for describing the details. There are three types of Data Set: Wide, Long, and Dimensional. The descriptions of these are very similar, differing mainly in the way they identify specific data cells.

Variables are described separately from the Data Structure, because variables can be re-used across many different structures. I could, for example, have a 'Nation' variable which is always represented with a two-character ISO Country Code. This variable might be reused across a wide body of different data sets (think of data in the SDG Indicators and the Minimum Set of Gender Indicators - both have such a Country variable, and share some others, but are mostly different structures). The role of a variable within a specific Data Structure is represented by the cdi:Component class. There are several specific types of Components:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

identifiers, measures, attributes, and so on. These are described as appropriate for each of the three types of data structure.

Most data sets can be described as Wide Data Sets in DDI-CDI – these are those cases where each row in a tabular view of the data is a record, and each column is a single variable. (When in doubt, this is the place to start in most cases.) Long Data Sets are often seen in cases where there is a stream of data, and the measured values are accompanied by another field which tells you which type of measurement they are. This case is important, although probably less common. Dimensional Data Sets are 'cubes' where the data points are addressed with the coordinates of a matrix, along different dimensions. This is a very common case for aggregated data.

6.6.5.1.3.1. Wide Data Structures

To describe a Wide Data Structure, each variable (here we use Instance Variables, acting in their capacity as Represented Variables) is assigned as defining a specific type of Component. In Wide Data Structures, the only Component classes we need are Identifier Components (cdi:IdentifierComponent), Measure Components (cdi:MeasureComponent), and Attribute Components (cdi:AttributeComponent). The diagram below shows how these and other classes are related.

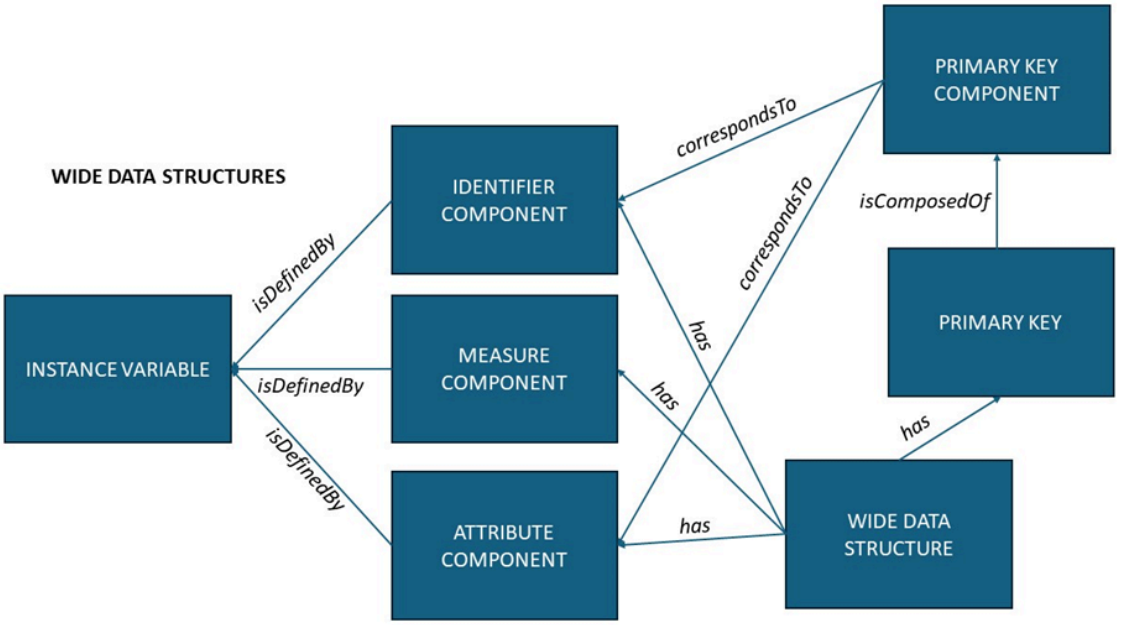

Figure 16. Wide data structures.

Measure Components hold the observed or measured values in our data. Identifier Components hold specific identifiers assigned to the cases in our data: each record is about a single 'case', and these are sometimes assigned identifiers to distinguish them.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

As in ***the example provided to the right***, this could be a tax-payer ID for a person, for example, or might be a randomly assigned number in the case where the cases have been anonymised to protect their identities.

Taxpayer ID	Year	Taxes Paid
003-49-6574	2001	2400.00
003-50-1234	2011	3600.00
009-42-6324	2021	2100.00

It is sometimes the case that an identifier Component is sufficient to disambiguate all the records in the data set: Identifier indicating the row/case, and the variable telling you which value in that record. Often, however, there will be additional fields which are needed (like a timestamp, in the case where the same case was measured/observed on more than one occasion). Attribute Components contain additional information about the measurements/observations which are applicable to each case. Typical Attribute Components are time stamps, geographical coordinates, statuses, or other data related to the context of the measurement or observation.

The Primary Key is the set of Components needed to uniquely identify each record in the data set. It is not typically the case that the Measure Components form part of the Primary Key.

Long vs. Wide Data Sets: If a Measure Component in a data set cannot be clearly associated with a single variable, but relies on the value of another variable to indicate which variable it measures, then you have a Long Data Set, and not a Wide one.

To illustrate the difference, some examples may be helpful. The table below is a simple example of a typical Wide Data Set:

Here, the Taxpayer ID is an Identifier Component (it gives the unique ID of each case), the Taxes Paid column is a Measure Component, and the Year is an attribute. If the cases are restricted to one entry per taxpayer, then the Primary Key will need only the Taxpayer ID field, but if (as is likely, given the need to pay taxes annually) there was more than one entry per taxpayer, then the Primary Key would also need to contain the Attribute Component Year. Compare this to the example in the *Long Data Structures* section below.

6.6.5.1.3.2. Long Data Structures

Long Data Sets are typically coming from sensors, administrative registers, or systems which contain 'spell data'. These forms of data are structured differently than what we see in most SQL-based systems, and they have a different relationship to the logical contents of data.

A second example can help to illustrate the difference between Wide and Long Data Sets. Long Data Sets use additional Components, as they have a more complex relationship to logical variables.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The table to the right shows a simple Long Data Set. Here, we have an Identifier Component (Patient ID), a column giving us the test performed (Test), a column with a measurement (Reading), and a column with a unit of measure for the reading (UoM). The UoM column is an Attribute Component: each entry provides a value that represents a Unit of Measure, and we can easily describe a 'Unit of Measure' variable (even though that can be inherent in the definition of a variable as is often seen in data sets today.)

Patient ID	Test	Reading	UoM
GX6574	Pulse	80	BPM
GX6574	Temp	37.00	Celsius
GX6574	Weight	75.00	KG

The Reading column is more problematic: each cell contains a measurement, but they are different types of measures. In order to understand the measurement, you have to consult the Test column, which tells you whether it is a pulse, a temperature, or a weight. This dependency is the characteristic feature of Long Data Sets.

When we describe a Long Data Structure, it is important to distinguish between the *logical* content and the presentation of it. Consider this re-structuring of the information shown in ***the example to the right***.

Patient ID	Pulse	Weight	Temp
GX6574	80	75.00	37.00

The values in the Test column have been broken out into separate columns which cleanly align with the measurements being taken, containing the values formerly held in the Reading column. The UoM column is now part of the definition of each of these variables, and so does not need to be included here. Logically, however, this is the same data as that shown above. (This re-structuring is illustrative – we are not recommending that Long Data Sets be re-structured as Wide Data Sets for publication.)

In order to describe my Long Data Structure, I need to define the *logical* variables, as seen in this re-working of the table, so that each variable can have a single, 'clean' definition, containing a single type of measurement. Each becomes an Instance Variable in my metadata. I would then go back and identify the additional variables needed to support the original presentation of the data: the Test column, the Reading column, and the UoM column.

UoM would be described as an Instance Variable, but Test becomes a special type: a Descriptor Variable (cdi:DescriptorVariable), and Reading becomes a Reference Variable (cdi:ReferenceVariable), which is the name for a variable which combines measurement values from several logical Instance Variables. These new types of presentational variables have different component types. The UoM column is still an Attribute Component, but the other columns introduce new ones: a Variable Descriptor Component

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

(cdi:VariableDescriptorComponent), which is defined by a Descriptor Variable (Test), and a Variable Value Component (cdi:VariableValueComponent), which is defined by a Reference Variable (Reading).

Thus, my resulting metadata will have descriptions of both my logical (Patient ID, Pulse, Weight, Temperature) and presentational (Test, Reading, UoM) variables. The Components in my Long Data Structure would reference the presented variables: Patient ID (an Identifier Component), Test (a VariableDescriptorComponent), and Reading (a VariableValueComponent) but the logical variables in turn would all be referenced from the VariableDescriptorComponent and associated classes, as described below.

VariableDescriptorComponents have a *refersTo* association with a VariableValueComponent (in this case, Test will associate itself with Reading in this way).

Descriptor Variables – the only variables which act as Variable Descriptor Components – do not have the same type of representations as other variables. Instead, they will always take their values from a Descriptor Value Domain (cdi:DescriptorValueDomain). This is a class which associates with a set of Descriptors (cdi:Descriptor), which are codes that refer to logical Instance Variables using an *identifies* relationship. The code in the Descriptor Component indicates which Instance Variable the value provided in the corresponding Reference Variable should be associated with. Thus, in our example above, when the :Test" column has a value of 'Temp', we know that the value in the "Reading" field is a measurement of temperature, associated with the temperature *logical* variable.

Reference Variables always have a representation which is the super-set of the values of the variables for which they can hold measures. They are thus often declared as generic types such as XSD string. The specific types of any value can be determined from the description of the associated Instance Variable (the logical variable) and do not rely on the Reference Variable. Figure 17 shows how these new presentational classes are connected.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

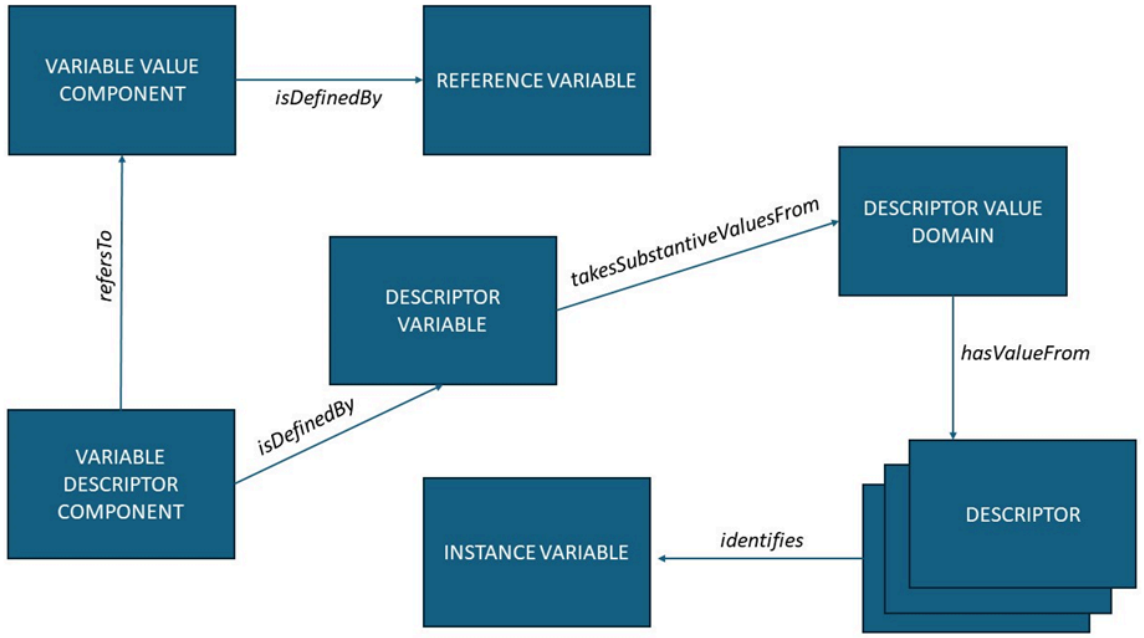

Figure 17. Long data structures - presentational elements.

The Primary key will be assembled from the *presentational* variables, however: in this case, a combination of the Patient ID and the Test.

The point of having this additional structural metadata for a Long Data Structure is that by providing logical Instance Variables, we are able to reassemble the values in the data set according to other structural arrangements (typically Wide or Dimensional), and to do so programmatically. Although we could have described the example table as a Wide Data Set, the actual variables in the data could not then be re-arranged or re-used: they would be specific to the structure of the data set they appeared in.

6.6.5.1.3.3. Dimensional Data Structures

Dimensional Data Sets have a Dimensional Data Structure which is similar to the Wide Data Structure, but has a different kind of Primary Key. The diagram below shows the set of classes which are needed.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

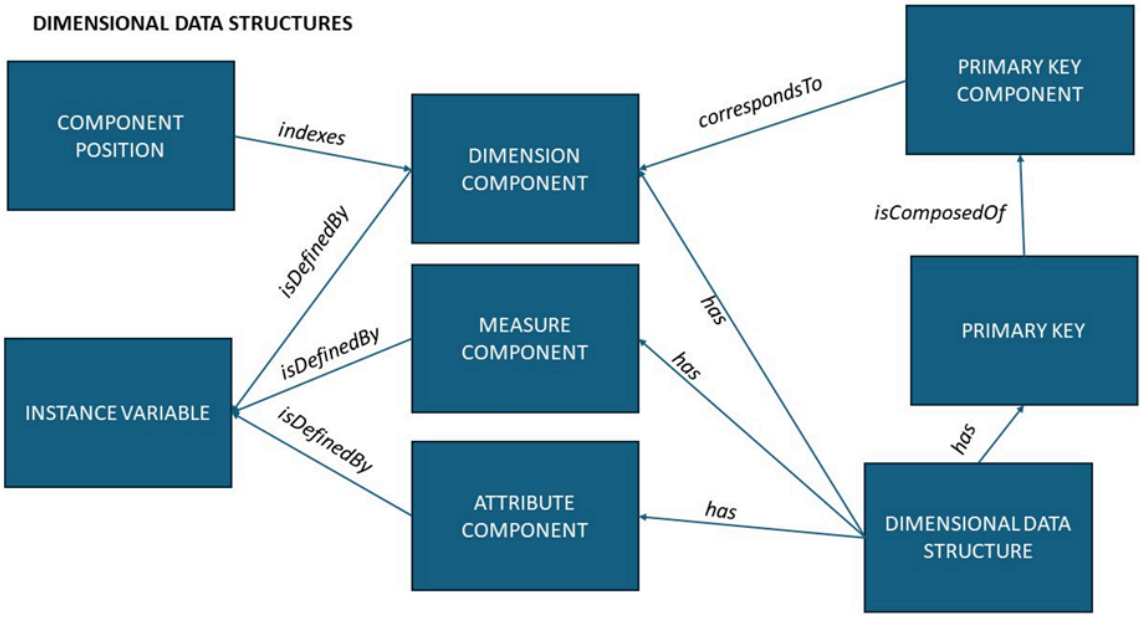

Figure 18. Multidimensional data structures.

In describing a Dimensional Data Structure, our Primary Key is only made up of Dimension Components. Taken together, these address not an entire record, but an individual cell with a cube – a multi-dimensional matrix. That cell holds the value of the Measure Component. Attribute Components are associated with the Measures at the cell level, not at the level of an entire record.

The exception to this is if a 'cell' has more than one Measure Component, in which case the set of Measure Components requires further disambiguation. This can be specified by qualifying the Primary Key values with the Measure Component/Variable. (It is recommended that only a single Measure Component be used if practical — cells should, if possible, hold simple values such as a string/code, a date-time, or a number.)

In some cases — notably in the DataCube Vocabulary from W3C and the SDMX specifications on which it is based — the dimensions in a key are ordered. In DDI-CDI, specifying an order is done using the Component Position (cdi:ComponentPosition) class. (In fact, you are always allowed to use cdi:ComponentPosition for any type of Data Structure, but in this profile we only recommend it for Dimensional Data.)

The table, 'Education Level of Belgian Residents', provides a simple example of a Dimensional Data Set, and how its structure can be described.

Education Level of Belgian Residents

Year	Degree	Province	Age	Percent
1986	BA	Antwerp	21-30	.25
1986	PhD	Brabant	31-40	.10
1987	BA	Limburg	21-30	.36

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

In this example, the first four columns act as dimensions for addressing the value given in the last column. There would be five Instance Variables described: Year, Degree, Province, Age, and Percent. The representations of these variables would be enumerated using SKOS Concept Schemes (see above) except for Year, which would be an XSD gYear, and the Percent variable, which would have a numeric type.

Year, Degree, Province, and Age would all be used to define Dimension Components, which could be indexed using instances of the Component Position class, assigning them a ranking from 1 to 4 (ordered low-to-high). Taken together, these four Dimension Components would form the Primary Key. Percent would be a Measure Component.

6.6.5.1.3.4. Notes on Data Structures and reuse

It is evident that many different types of Components are shared between different Data Structures. It is important to understand that although this is true – and that any given set of data may be organised according to more than one type of Data Structure, the instances of Components themselves are *not reusable* by definition. A Component is the specific role played by a variable in the context of a data structure. Each Component appearing in a Data Structure is specific to that Data Structure, and should have its own identity. The variables are the reusable form of the information in describing the same thing used across different structures.

Remember, too, that while the DDI-CDI model uses Represented Variables – the reusable part of a variable description – in modelling Data Structures, for the purposes of CDIF we rely on the fact that Instance Variables are a subclass of Represented Variables, and can thus be used *as* Represented Variables. (This is expressed with the rdf:instanceOf relationship on the instances of cdi:InstanceVariable – see Appendix 4, below.)

6.6.5.1.3.5. Describe the physical format of the data

DDI-CDI separates the logical and physical organisation of data. The way that data is stored in a system or recorded in a file has no necessary relationship to the way in which the data is logically structured.

In this initial version of CDIF, we are only describing data that is expressed in a textual format such as a CSV file, fixed-width tabular files, or the query result from a relational system. (This is a limitation which will be addressed moving forward.) Further, the assumption is made that any Data Set uses only one structure for all of its records, and that these are organised in a uniform way.

These recommendations pertain to data that can be represented in a tabular data structure, packaged in datasets that consist of a set of records that all have the same set of fields. Figure 19 shows the classes which are needed for describing the physical encoding of the data.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

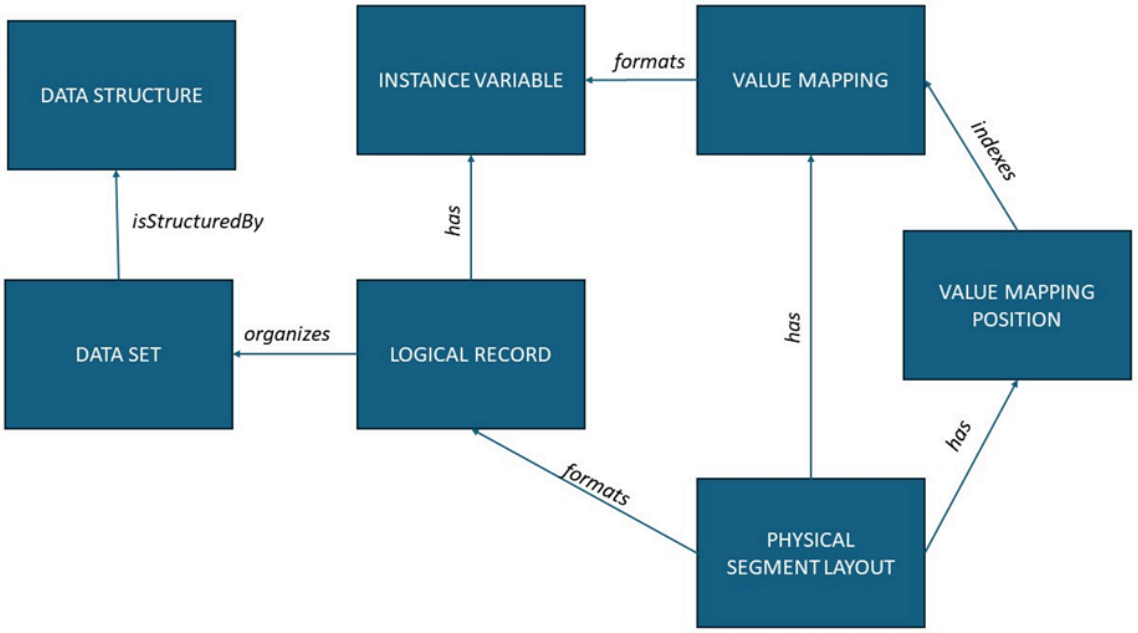

Figure 19. Describing physical layouts.

The Logical Record (cdi:LogicalRecord) references all of the Instance variables used within the Data Set. A Physical Segment Layout (cdi:PhysicalSegementLayout) describes the way in which that Logical Record is expressed in the physical file. It has a set of Value Mappings (cdi:valueMapping) instances and corresponding Value Mapping Position (cdi:ValueMappingPosition) instances, which provide the links between the physical layout and the values of the Instance Variables. Physical Segment Layout instances contain much of the information needed by machines to read the data (i.e., character encodings, delimiters, line-end characters, etc.).

(Note that the *formats* relationship between Value Mapping and Instance Variable has been collapsed from what is presented in the DDI-CDI model by omitting the intervening Data Point, as Data Points are not instantiated in this profile.)

6.6.6. Describing the integration of data sets

This version of the CDIF recommendations does not contain a full profile for the description of data integration, but this topic has received a lot of attention from the working group. An exploration into different data integration scenarios has been conducted, notably an effort to integrate data from ILO, the World Health Organization (WHO), and the SDG Indicators, with a goal of publishing this data into Google Data Commons and the knowledge graph model used there. This work is on-going, but it has demonstrated the set of metadata needed to fully describe a data integration.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

This set of metadata requirements can be summarised as follows, based on the assumption that there is access to the data being integrated:

1. Detailed data description (variable-level)
2. Structural metadata of the data sets
3. Enumerated values (codelists and classifications)
4. Mappings between sets of enumerated values
5. Processing description, indicating how mappings were implemented in transformations, and what other operations were performed for data integration

While this seems like a daunting set of information, the exploratory work has shown that if the metadata are available in a sufficiently detailed form, then the actual integration itself is straightforward.

6.7. Universals: Time, Geography, and Units of Measurement

Some information that is not domain-specific is commonly required to describe and understand resources. The use of common patterns and encodings for this information helps cross-domain interoperability. We can think of the term 'universals' as that metadata covering properties that are inherent in many different kinds of measurements, across many different disciplines and domains. Often, these are important in describing aspects of data which are key to the integration of data from different sources.

As a general principle, for any **value** in a CDIF metadata description it should be clear what **value space** it is taken from (i.e. scale, reference-system, or vocabulary), and this should be traceable to a suitable definition and authority. Where possible CDIF shall recommend a default approach for universals, as well as a small number of variants to accommodate commonly encountered applications.

We provide recommendations for three specific universals: space, time, and units-of-measurement (UOM). Note that practice for universals varies across disciplines according to their different requirements. For spatial data in particular, a single representation is not possible.

UOM are in some cases complex, and there is on-going discussion about how best to represent the more complex cases. In CDIF, we do not address the description of quantities at this level - we would defer to work in groups such as CODATA's Digital Representation of Units of Measure (DRUM) group.[140] This topic will be addressed more completely in future by CDIF as more clarity emerges.

6.7.1. General pattern for implementation of universals

The representation of universals is always implemented in some frame of reference:
(value, reference-system [, time]). The (optional) 'time' value is to allow for reference-systems subject to change, where the interpretation of the value may depend on a particular version of the reference-system.

[140] https://codata.org/initiatives/task-groups/drum/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

In some cases, the structure of data may provide these values in an explicit way which does not require the use of patterned text (e.g., there are separate fields for each piece of the information.) When available, these should be utilised in preference to patterned text. Fields in data might contain values defined by the data schema to use a particular reference system, or each value field might be paired with a field that assigns the reference system for that value.

6.7.2. Geography

In order to support discovery and selection of datasets, we need to:

1. describe the spatial extent or footprint of a dataset *(e.g. name, bounding box)*
2. Say something about the spatial distribution and representation of values within a dataset *(e.g. grid definition, point-cloud, precision, spacing)*

Where not specifically indicated in CDIF, location should be expressed according to the pattern (value , location-system or authority [, time]), for example:

- (coordinate geometry, Coordinate Reference System (, date))
- (placename, Gazetteer (, date))
- (cell id, grid definition)

6.7.2.1. Location systems

There are many different location systems, including:

- Coordinate-based location, used for
    - Point location
    - Bounding box
    - Polygon
- Addresses, delivery points, lot numbers
- Named places, points-of-interest
- Named areas:
    - Administrative areas (many ranks and functions)
    - Post-codes
    - Electoral districts
    - Statistical areas
- Grids, DGGS

Different applications use different systems, reflecting information requirements which often cannot be controlled by those managing or disseminating data (e.g., in the natural sciences you may find coordinates or grid-cells, where dissemination and implementation science might use coordinates or point-of-interest, social scientists and official statisticians would use named areas, statistical areas, or administrative areas, and utilities might use addresses or post-codes).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

It is often best to describe locations using the systems employed by the creator of the data, leaving it up to the data integrator to perform needed translations, because the approach to harmonising locations can be driven by the methods employed for analysis and depend on the research question in a particular case. What we can do, however, is to make sure that needed information is present to unambiguously understand the locations as described. The following sections look at some of the common location systems in use.

6.7.2.2. Coordinate-based location

6.7.2.2.1. Coordinate values

For points, lines, and polygons (including bounding-boxes) coordinate-based locations are used - usually longitude and latitude, in decimal-degrees.

- It is important to pay attention to the **coordinate sign**: longitudes in the western hemisphere (west of the Greenwich meridian) are negative, and latitudes south of the equator are negative.
- **Coordinate order** is also important: in traditional cartographic and navigation systems the order was (latitude,longitude), i.e. (y,x); however most modern digital systems follow the more common (x,y) order from maths and graphics i.e. (longitude, latitude).

With unfamiliar data, it is always worth checking both sign and coordinate order before doing any other data manipulations.

6.7.2.2.2. Coordinate reference system

The most commonly used coordinate reference system (CRS) for geographic coordinates is WGS 84.[141] This is the default system used by GPS receivers, and most Web-mapping applications, so if the CRS is not explicit it can generally be assumed to be WGS 84. This is also the default for the common geometry representations, such as GeoJSON[142] and the Well Known Text Representation of Geometry (WKT)[143].

The definition of a CRS involves several pieces of information, including the axis directions and measurement units, the origin location where the value is (0,0), as well as the ellipsoid that is used to approximate the shape of the earth. Some CRS are three-dimensional, including elevation. The de facto authority for definitions of coordinate reference systems is EPSG[144]. They maintain a registry with a Web API at https://apps.epsg.org/api/swagger/ui/index. The EPSG is used by many mapping software systems, and other Web copies are available (e.g. https://epsg.io). WGS 84 is denoted EPSG:4326 and the full definition can be found at https://apps.epsg.org/api/v1/CoordRefSystem/4326/export/?format=wkt.

[141] https://earth-info.nga.mil/index.php?dir=wgs84&action=wgs84
[142] https://en.wikipedia.org/wiki/GeoJSON
[143] https://en.wikipedia.org/wiki/Well-known_text_representation_of_geometry
[144] https://epsg.org. Properly speaking, EPSG now refers to a set of products maintained by the International Association of Oil and Gas Producers (IOGP)'s Geoinformatics Committee. These products originated with the European Petroleum Survey Group (EPSG), which merged with the IOGP in 2005.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Some jurisdictions have legal requirements to use other coordinate reference systems for some applications for official purposes. For example, the British National Grid is standard in the UK, in which the coordinates are given in metres, and the origin is fixed so that all locations within the UK have positive coordinate values. The British National Grid is denoted EPSG:27700 and the full definition is found at https://epsg.org/crs_27700/OSGB36-British-National-Grid.html or https://epsg.io/27700.

6.7.2.3. Shapes

Point-location is defined by a single set of coordinates. Areas are usually represented by their perimeter polygon. This is defined by an ordered sequence of points, where the last point should coincide with the first in order to close the polygon. The minimum and maximum values for an extensive location define a bounding-box. The orientation of the boundaries of a bounding box align to the CRS used to describe the corners of the box. For example, for coordinates specified using latitude and longitude, the bounding box is delimited by meridians and parallels.

GeoJSON and WKT provide specific encodings. DCAT provides the following predicates that relate spatial information to a dataset:

1. `dcterms:spatial` for the spatial extent of the data;
2. `dcat:spatialResolutionInMeters` for the spatial precision or spacing of items within a dataset.

The Global Biodiversity Information Facility (GBIF) has a geo-referencing best practice guide that covers latitude, longitude, altitude, and depth. This is a good resource for determining good practice.[145]

6.7.2.4. Identifier-based location

Nominal systems associate a *name* or *code* to some spatial location or region. Commonly used cases include:

- Countries;
- Administrative units (e.g. NUTS - Nomenclature of Territorial Units for Statistics[146]): states, provinces, cities, local government areas);
- Statistical areas (defined by statistical agencies or census bureaus, e.g., the Australian Statistical Geography[147]);
- Electoral districts (whose geographic footprint may change frequently so date is required);
- Postcodes.

[145] https://doi.org/10.15468/doc-gg7h-s853
[146] https://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics
[147] https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026#asgs-diagram

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.7.2.4.1. Reference system

In most - but not all - cases, there is a well-defined mapping of the name or code to a geospatial area. For *names* the list of mappings is called a *gazetteer*. The authority for this mapping may be well known, or may be more informal.

- Statutory naming authorities typically have a formal process for gazetting (publishing) geographic names.
- Postcodes are usually well-known and often convenient, however the exact mapping to space may be proprietary to a local postal delivery service.
- GeoNames[148] is a crowd-sourced service that associates a point-location with many geographic names. The location used for arealy-extensive places can be inconsistent.

In many or most cases the mapping of a name to a geospatial area is time dependent. In some cases the area attached to a name is contested, for political and cultural reasons, or because of historical uncertainty. Contemporary name-based systems are generally managed nationally or locally. Historical systems are generally more difficult and there may not be an authoritative source. Geonames is a pretty good general-purpose service, but it only gives a point-location falling somewhere within the place.

6.7.2.5. Grids

Where a dataset is composed of variables or properties whose value varies across a spatial domain, it is often represented as sampled discretely at locations on a regular grid. Discrete Global Grids (DGG) are an emerging alternative to cartesian grids.

6.7.2.5.1. Reference system

The most common grid arrangement is specified in terms of an origin and axis directions (with respect to a CRS), and cell spacing. A location within the grid is then indicated by an integer index or count e.g. (234, 8916). Spatial grids may be one-, two-, three-, or four-, (spatio-temporal) dimensional.

6.7.2.5.2. Usage

Spatial variation of a property across a spatial domain is commonly represented and exchanged using grids. Visualisations almost always use a cartesian (rectangular) gridded representation. Grids are the most common representation for spatial analysis and numerical modelling. Grids are convenient for data integration. However, they may need to be re-sampled or interpolated to a common orientation and spacing. Different DGGs are efficient for topology and proximity analysis, and area-based calculations.

6.7.2.6. Addresses

Many different sources exist for the description of addresses. Among these are EPSG, ISO 19101-1:2014[149] and ISO 19107:2003[150].

[148] https://www.geonames.org/v3/
[149] https://www.iso.org/standard/59164.html
[150] https://www.iso.org/obp/ui/#iso:std:iso:19107:ed-1:v1:en

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.7.3. Time

In order to support discovery and selection, we need to:

1. Describe the time period covered by a dataset (e.g., time-interval, named period, reference period)
2. Say something about the temporal representation within a dataset (e.g., precision, spacing, relative times or sequence/ordering, frequency of time series)
3. Provide information related to the time a dataset was prepared, updated, and its period of validity

DCAT provides the following predicates that relate a time to the dataset:

1. `dcterms:temporal` for the temporal extent of the data.
2. `dcat:temporalResolution` for the temporal precision or spacing of items within a dataset.
3. `dcterms:issued`, `dcterms:modified`, `dcterms:accrualPeriodicity` relate to the dataset management.

The description of time will use some representation within a *temporal reference system* (TRS). The TRS must be known in order to understand the value. There are several representations of time in common use[151].

6.7.3.1. Temporal coordinates

Time can be understood as involving positions along the timeline. Position can be expressed as a *coordinate* in a one-dimensional system. Some practical time systems use this approach explicitly:

- Unix[152] and GPS time are based on seconds counted from an origin in 1970 and 1980, respectively.
    - GPS time is actually represented using a pair of numbers for week number plus seconds into week (optimised for low-bandwidth communication).
- Julian day counts the number of days since the beginning of 4713 BCE.
- Ordinal date counts the number of days from the beginning of a year.
- Geological applications express time in years, or millions of years Before Present (BP)[153], which is defined as 1950 for applications where the precision of the measurement technique warrants.

Note the differing precision of these systems: seconds, days, years, millions of years. The origin of a temporal coordinate system is called the *Epoch*.

6.7.3.2. Calendar and clock

Both calendar and clock represent a time position as a set of integer or decimal values for nested elements of progressively finer resolution: year, month, day, hours, minutes, seconds. ISO 8601[154] provides a textual representation or microformat which has been adopted or implemented in most encoding and software systems.

---

[151] Open Geospatial Consortium Abstract Specification - TOPIC 25 - ABSTRACT CONCEPTUAL MODEL FOR TIME - OGC doc 23-049 provides a conceptual framework for these different representations - (to be published) - draft https://portal.ogc.org/files/?artifact_id=107087&version=2

[152] https://en.wikipedia.org/wiki/Unix_time

[153] https://en.wikipedia.org/wiki/Before_Present

[154] https://en.wikipedia.org/wiki/ISO_8601

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Example: “2023-10-04T12:13:30.25+02:00”

This should be understood as a multi-valued *tuple* providing values for each component of the calendar and clock. Conversion to a coordinate on a time-line can be done arithmetically *except* that a look-up is required to convert months to their durations in days or seconds, etc. The final two elements of the value, following ‘+’ or ‘-’, indicate the timezone, encoded as the hours and minutes offset from UTC[155]. ISO 8601 uses the Gregorian calendar, whose epoch is the beginning of Year 1 CE.

6.7.3.2.1. Usage

While other calendars are in contemporary use within some specific communities and for some cultural applications (e.g. Traditional Chinese, Julian, Islamic, Jewish, Baha’i), the Gregorian calendar and the 24-hour clock are universally used for technical purposes.

6.7.3.3. Ordered nominal timescales

In historical, archeological, and geological applications, dates may be expressed using a named period, tied to a recognised culture (e.g., ‘bronze age’), dynasty (e.g., ‘Tudor’), or defined by some event(s) observed in the geological record tied to the natural history of a region or the earth (e.g., ‘Proterozoic’). Some of these systems can be mapped to temporal coordinates, though calibrations may be adjusted according to new evidence[156]. In general, however, only the **ordering relationships** between members of a nominal timescale are defined - i.e. we know that a nominal date is before, after, or during some other date within the same system, but not the size of the separation between them.

There are many nominal systems. The International Chronostratigraphic Chart[157] and its calibration is authoritatively maintained by the International Commission on Stratigraphy, and provides the temporal reference system used internationally in geology. Most other nominal systems are only used locally, within particular disciplinary communities.

6.7.3.4. Recurring and periodic times

Many datasets are structured as time-series, with either regular or irregular sampling period, or frequency of observation. A sequence of time-stamps may be thought of as a one-dimensional grid.

6.7.3.4.1. Usage

Time series - the periodic measurement of the same phenomenon, using the same methods, etc. - are a common way of determining trends or otherwise making observations across time. In some cases, time series are the product of data integration from disparate sources, while in other cases the data are collected as part of time series as part of their design. In the Statistical Data and Metadata Exchange (SDMX) *Content-Oriented Guidelines* there is a standard enumeration of commonly used frequencies for national and

[155] https://en.wikipedia.org/wiki/UTC_offset
[156] Calibration of the geologic timescale is the primary activity of the International Commission on Stratigraphy.
[157] https://stratigraphy.org/chart

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

supra-national official statistics.[158] ISO 8601-1:2019[159] adds notation for recurring arbitrary intervals (e.g., 9-5, every Mon-Fri, etc.).

6.7.4. Units of measurement

Quantitative data is always expressed using a scale, commonly referred to as the *unit of measurement* (UOM). The representation of any quantity can be conceived as a 'tuple' comprised of:

- a number
- the scale or UOM
- (optionally) a quantity-kind, which provides context and semantics for the quantity

The quantity-kind is sometimes required because the same unit of measure can be used for incompatible quantity-kinds (e.g., energy and torque can both be quantified as Newton-metres). When encoding quantities, there are three common patterns:

- Micro-format, using a standard uom notation, and separator between the number and UOM:

```
109.5 km/hr
```

- Tuple or data-structure:

```
{
"@value" : "109.5" ,
"@type" : "https://si-digital-framework.org/SI/units/kilometre.hour-1"
}
```

- Fix the UOM for an array or collection, where a header or metadata gives the uom for all values, and the datastream then provides an array of values only:

```
( “km/hr” , ( 60.4 , 75.1, 99.0, 109.5 ) )
```

For any of these approaches the code or symbol that denotes the uom must come from a well defined system with unambiguous semantics.

Note that CDIF Data Description provides properties on variables for specifying both UOM and quantity-kind (see Appendix 4), using the micro-format described in Appendix 1.

[158] https://sdmx.org/?page_id=3215

[159] https://www.iso.org/standard/70907.html#:~:text=This%20document%20specifies%20representations%20of,Coordinated%20Universal%20Time%20(UTC)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

6.7.4.1. UOM code systems

The following list is provided for reference, giving information about some of the common coding systems used for UOM.

SI Digital Framework - https://si-digital-framework.org/

- Scope: Semantic reference for SI base and derived units
- Maintained by BIPM https://www.bipm.org/en/ - the official custodian of the *International System of Units* (SI) https://en.wikipedia.org/wiki/International_System_of_Units
- Symbols for units follow ISO/IEC 80000 (incl special characters, superscripts etc)
    - *m, eV, h, km/h*
- Each UOM is denoted by a URI e.g.
    - SI Units and Quantities
        - Unit https://si-digital-framework.org/SI/units/metre
            - Quantity https://si-digital-framework.org/quantities/LENG
        - Unit https://si-digital-framework.org/SI/units/pascal
            - Quantity https://si-digital-framework.org/quantities/PRES
    - Non-SI units accepted for use with the SI units
        - Unit https://si-digital-framework.org/SI/units/electronvolt
            - Quantity https://si-digital-framework.org/quantities/ENGY
        - Unit https://si-digital-framework.org/SI/units/hour
            - Quantity https://si-digital-framework.org/quantities/TIME
    - Prefixes
        - Prefix https://si-digital-framework.org/SI/prefixes/kilo
        - Prefix https://si-digital-framework.org/SI/prefixes/nano
    - URIs for compound units based on SI Units are formulated as follows:
        - Unit https://si-digital-framework.org/SI/units/kilometre.hour-1

QUDT (Quantities, Units, Dimensions and Types) - qudt.org

- Scope: semantic descriptions of UOM for science and engineering
- Enumeration of ca. 2500 UOM - qudt.org/doc/DOC_VOCAB-UNITS.html
- Each UOM is denoted by a URI - e.g.
    - http://qudt.org/vocab/unit/M
        - http://qudt.org/vocab/quantitykind/Distance
        - http://qudt.org/vocab/quantitykind/Length
    - http://qudt.org/vocab/unit/PA
        - http://qudt.org/vocab/quantitykind/Pressure
        - http://qudt.org/vocab/quantitykind/Stress
    - http://qudt.org/vocab/unit/EV
        - http://qudt.org/vocab/quantitykind/Energy
    - http://qudt.org/vocab/unit/HR
        - http://qudt.org/vocab/quantitykind/Time

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

    - http://qudt.org/vocab/prefix/Kilo
    - http://qudt.org/vocab/prefix/Nano
    - http://qudt.org/vocab/unit/KiloM-PER-HR
- Local-name defined by a rule (non-opaque, almost computable)
- Dereferencing URI gets a definition of the UOM, with link to dimensionality
    - Enough information to support unit conversion
    - Codes and symbols from various systems, including UCUM, SI, and UN/ECE Recommendation 20[160]
- System also has an enumeration of 'quantity-kinds' - i.e. semantics
    - E.g., energy and torque have the same scale and dimensionality, but should not be transformed
- Maintained by volunteers
- Change requests processed ca. weekly - https://github.com/qudt/qudt-public-repo/issues.

UCUM (Unified Code for Units of Measure) - https://ucum.org/

- Scope: codes for UOM for science
- Terminal symbols plus a rule for constructing arbitrary units - extensible/scalable
- Codes match traditional scientific notation, with simplifications to match 7-bit ascii keyboard (no special characters, italics, superscripts) - e.g. km/hr, km.hr-1, N.m-2
- Maintained by LOINC[161] - Biomedical/clinical standards authority
- Change requests processed slowly - https://github.com/ucum-org/ucum/issues
- Public specification document at https://ucum.org/ucum
- Validation and conversion
    - interactive https://ucum.nlm.nih.gov/ucum-lhc/demo.html
    - API https://ucum.nlm.nih.gov/ucum-service.html.

6.8. Implementing CDIF Core Profiles

It is possible to implement the CDIF profiles singly or in combination when describing a FAIR resource. It is important that those reading the provided metadata know what fields are going to present, and what profiles are being employed. To indicate this, CDIF uses the *dcterms:conformsTo* property, with the following values:

CDIF_core_discovery_1.0

CDIF_core_access_1.0

CDIF_core_data_description_1.0

CDIF_core_controlled_vocabulary_1.0

[160] https://unece.org/trade/documents/2021/06/uncefact-rec20-0
[161] https://loinc.org/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

(There is some discussion as to whether URI tokens would be more useful, as the host might change, but the current recommendation is as stated.)

In many cases, there is little or no overlap between the profiles listed above, but there are some special cases where the combined use of profiles demands that some additions be made to what is indicated elsewhere. In the case where both the Core Discovery profile and the Core Data Description profile are used together, there can be two descriptions of variables. The Discovery profile uses Schema.org Statistical Variables, and the Data Description uses DDI-CDI variables. In this case, the variables should be described only once, with the following additions:

- Each variable should declare that it is both a Schema.org Statistical Variable and whatever type of DDI-CDI variable it is.
- The *name* property should be typed as both a Schema.org *name* and a DDI-CDI *name.*
- The Schema.org *description* should be duplicated and separately typed as a DDI-CDI *definition* property (or vice-versa). These two properties are similar but not identical, and may (or may not have) different content. In the case where only a single description/definition exists, it should be duplicated. Note that DDI-CDI ReferenceVariables do not have this field, and can be treated as an exception by omitting the declaration of the DDI-CDI *definition* property.

Note that when the Core Data Description is used and includes the description of enumerations (as opposed to simply referencing SKOS Concept Schemes from elsewhere), the Core Controlled Vocabulary profile should also be stated in a *dcterms:conformsTo* declaration, as both profiles will be used.

7. Future CDIF profiles and topics for further work

This section describes some areas of future work for CDIF. Some of these topics have not been points of focus during the WorldFAIR project, while others have. In each of these cases, it is simply not yet possible for CDIF to recommend a particular practice, because of the current state of the art and the need for further exploration of potential commonality of practice across domains. For each topic, the current status will be described.

7.1. Provenance

7.1.1. What is “provenance”?

One major topic of discussion within the FAIR community - and mentioned in the FAIR Data Principles in a significant fashion - is provenance:

> “R 1.2 (Meta)data are associated with detailed provenance”

Provenance is a topic which can be understood in many different ways, and the FAIR Data Principles do not provide specifics on what is and is not classed as “provenance.” For the purposes of CDIF, however, a more complete definition is useful. For this, we can turn to the W3C’s PROV Family of Documents, which reflects a

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

common understanding of the meaning at least for use on the Web. The PROV Overview[162] states: "Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness."

Eight areas of support are identified:

1. the core concepts of identifying an object, attributing the object to person or entity, and representing processing steps;
2. accessing provenance-related information expressed in other standards;
3. accessing provenance;
4. the provenance of provenance;
5. reproducibility;
6. versioning;
7. representing procedures;
8. and representing derivation.

The PROV Data Model[163] provides the following:

> **"provenance** is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing. In particular, the provenance of information is crucial in deciding whether information is to be trusted, how it should be integrated with other diverse information sources, and how to give credit to its originators when reusing it. In an open and inclusive environment such as the Web, where users find information that is often contradictory or questionable, provenance can help those users to make trust judgments."

From these we can usefully characterise what we mean when we talk about "provenance":

- The goal is to support the assessment of the quality and trustworthiness, of data, the understanding of fitness-for-purpose, and the ability to give credit;
- Key to this is identifying of the agents or actors who have operated on the data or resource, and description of the processes or procedures they have performed on it;
- Distinguishing different versions or forms of data - the inputs and outputs of processes - is a necessary aspect of provenance.

Taken broadly, we could understand "provenance" to encompass all metadata descriptions, but this is not what is intended. CDIF views provenance information as an additional set of metadata meant to complement that used for the purposes of data description (both structural and semantic), data access, and data discovery. This separation seems to be aligned with the sense of the W3C Family of Documents as they use the term.

[162] See https://www.w3.org/TR/prov-overview/

[163] https://www.w3.org/TR/2013/REC-prov-dm-20130430/#dfn-provenance

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

In our view, provenance does contain detailed information not only about who created a FAIR resource for citation purposes, but also describes the activity of creating or transforming that resource. Process is a critical part of provenance description, because it specifies exactly what was done at a detailed level, and this is significant to downstream users.

7.1.2. Relevant standards and specifications

There are a number of different standards and specifications used to describe aspects of provenance as we have defined it. It is notable that PROV mentions the need to connect provenance information described using different standards. These standards and specifications can be broken out into three categories:

7.1.2.1. General models: The PROV Family of Documents

W3C's PROV Family of Documents gives us a useful point of reference for engaging with provenance information, from general modelling to specific implementation. Arguably, it is too general for the specific tasks we want to use it for in CDIF, although there is not yet agreement on this point. It is true that many other specifications have taken PROV as a starting point, and extended it with additional classes for specific purposes.

7.1.2.2. Processing descriptions

There are several different specifications for describing processes and processing, intended to be used for a range of purposes. Among these are PROV-O (the ontology from W3C, and part of the PROV Family of Documents), the Common Workflow language (CWL),[164] the Business Process Modelling and Notation (BPMN),[165] PROVOne,[166] DDI-CDI Process descriptions,[167] the Validation and Transformation Language (VTL),[168] and the Structured Data Transformation Language (SDTL).[169]

These different specifications address different aspects of process description. CWL, BPMN, DDI-CDI, and PROVOne all look primarily at how workflows can be used to connect the actors and resources involved in a process, and can be used to connect other standard descriptions of things such as execution code. VTL, SDTL (and its latest un-released evolution, 'SDTH') are more focused on describing the actual execution of a process, rather than the workflow of which it is a part. Of these standards, PROV-O serves as a basic or underlying model for many of them, including PROVOne, DDI-CDI, and SDTH.

It should be noted that of these, only DDI-CDI explicitly addresses the case of 'declarative' process description, in which opaque 'black box' processes are invoked. This may be significant in describing the

[164] https://www.commonwl.org/
[165] https://www.bpmn.org/
[166] https://jenkins-1.dataone.org/jenkins/view/Documentation%20Projects/job/ProvONE-Documentation-trunk/ws/provenance/ProvONE/v1/provone.html
[167] https://ddi-cdi-resources.bitbucket.io/2023-11-12/field-level-documentation/DDICDILibrary/Classes/Process/index.html
[168] https://sdmx.org/?page_id=5096
[169] https://ddialliance.org/products/sdtl/1.0

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

operations of LLMs in future. CWL has been implemented in combination with RO-Crates[170] - one example of how many different standards can be used in combination.

7.1.2.3. Citational standards

There are several models for describing and attributing the roles involved in the development of citable assets and these are often combined with standards for the discovery of data and resources. This is perhaps a less urgent topic for CDIF than other aspects of provenance. It should be noted, however, that significant challenges remain in this area, notably around software citation, including the citation of algorithms which may not be released as citable, scholarly publications. This remains an area to be more fully explored.

One related area that is not directly classed as provenance, but which deserves mention, is quality standards. There are a number of quality frameworks of different types, and it may be important in future to consider how metrics and information about data consistency and quality might be significant in this area.

7.1.3. Describing the full data lineage chain

One requirement which has emerged from the discussions around provenance in large-scale FAIR exchanges is the need to gain visibility over the whole of the chain of data lineage. Most of the specifications described assume that a single organisation will be describing what they did with a set of inputs to produce their outputs. Typically, this is a research project taking or collecting data and metadata, processing it, analysing it, and documenting the provenance as part of the outputs, which would also include the data and other needed metadata.

In reality, the picture can be more complicated: the data used as an input for one project may itself be the product of a set of processes and analyses, based on other input data. If we are to understand the full lineage of data and other resources, we need to be able to describe all preceding steps, and - ideally - to add a provenance description of our own use of that data in further research activities. This requires that a series of players be able to pick up and extend the provenance descriptions provided by those who have interacted with the data in an “upstream” capability.

Today, this requirement is not addressed fully by existing standards, nor in practice. An interesting deliverable from EOSC Life, “EOSC-Life Common provenance model for processing biological material, data generation and computational workflows”[171] describes the issue, and proposes some approaches based on PROV, but it is clear that more work is needed.

7.1.4. Future provenance recommendations in the CDIF

At this time, CDIF makes no specific recommendations regarding provenance description. It is clear, however, that this area is critical in a full FAIR implementation for cross-domain use, and is directly connected to issues

[170] https://about.workflowhub.eu/Workflow-RO-Crate/

[171] Wittner, R., Geiger, J., Müller, H., Frexia, F., Mascia, C., Exter, K., Cancio, I., & Holub, P. (2023). EOSC-Life Common provenance model for processing biological material, data generation and computational workflows. Zenodo. https://doi.org/10.5281/zenodo.8279525

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

around data access, repositories, and replication. There are many different standards in these areas, and some alignment regarding how organisations and other actors are described may be necessary. As CDIF develops further, provenance is a topic which will be given a high priority.

7.2. Context

The information needed by the user of data in a cross-domain scenario is potentially different than what is typical of secondary data use within domains, Because researchers will be less familiar with the ways in which data are collected or produced, and potentially less familiar with the methods employed, it is sometimes necessary to formalise dependencies between different fields within the data which cannot be known from a strict structural or semantic description.

In CDIF, we use the term 'context' to indicate this additional information which is related to, but not the same as, what we cover in recommendations around provenance and data description.

A simple example of such a context is the need for a pair of measurements with a dependence between them. Take, for example, a blood pressure data set in which the systolic and diastolic readings are contained in separate fields within the data (as separate variables) with a reading taken for each patient while reclining and while standing. It might be useful to connect the two values for the reclining measurements, and the two values for the standing measurements, as each pair forms the typical systolic-over-diastolic expression of the blood pressure reading, and is in fact the value to which the unit of measure - 'mmHg' applies.

Two specifications which help us to understand this type of dependency (or 'context') are the I-ADOPT vocabulary[172] from RDA, and the OGC Abstract Specification Topic 20: Observations, Measurements and Samples standard which was stanard(jointly developed with ISO as ISO 19156:2023[173]

It may be possible to express the dependencies of such context in a descrip[tion of the data which employs DDI-CDI, using the cdi:VariableRelationship class, which provides a reference to an external source for the specific type of dependency. This approach was explored in the CDIF work, but no firm recommendations can yet be made.

7.3. Perspectives on AI

The use of artificial intelligence in science is a broad topic, and beyond the remit of CDIF. There are and will continue to be many innovative applications of AI in different forms in scientific analysis. CDIF restricts itself to discussing how FAIR implementation might impact the use of AI.

One view is that FAIR data will enhance the quality of training sets for large language models (LLMs), reducing the occurrence of 'hallucinations'. (This view came up in discussions at the recent Lorentz Center workshop 'Road to FAIR and Equitable Science'[174])

[172] https://i-adopt.github.io/
[173] https://www.ogc.org/standard/om/
[174] https://www.lorentzcenter.nl/the-road-to-fair-and-equitable-science.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

The release of Croissant, a metadata format for describing data sets for the purposes of training LLMs, indicates that the developers of these models are aware of the need for structural metadata in describing tabular data sets.[175] The similarities between the basic structural metadata provided by this format, and the variable-level description found in standards such as DDI-CDI are clear. While Croissant has less-complete information regarding the data, it is not intended as an integration or management format. It may make sense for CDIF to address this issue by publishing a mapping between its own recommendations for describing data in an 'integration-ready' form, and the Croissant format for LLM training sets. This remains to be investigated.

Another area where LLMs (and other AI applications) intersect with the concerts about FAIR implementation addressed by CDIF is in the area of provenance metadata. Traditionally, processes are described in a stepwise fashion, as we see in such standards as the Business Process Model and Notation (BPMN)[176]. LLMs and other AI applications which analyse data do not neatly fit into this model: they function as 'black boxes' within a process flow, with no clear causal connection between individual input data and the output model, performing operations in an opaque manner. It is clear that if we are to fully describe provenance, the incorporation of AI into a workflow description must be supported.

Similarly, material taken from the Web and used in training sets is not always of known provenance, and there are concerns regarding the ethical and legal correctness of the LLMs access to this data. Machine-actionable information regarding permitted access and use should make it clear to large-scale harvesters what is and is not permissible. The CDIF recommendations regarding access must take this requirement into account.

In general terms, we need to be aware of the 'AI readiness' of our data for these and other reasons: machines - and the LLMs are an increasingly important form of these - are becoming a major type of 'user'. This emphasises the importance of having machine-actionable metadata, which can reliably be retrieved and acted upon and with, in conjunction with our data.

It remains to be seen how generative computing will impact the availability of metadata, but early experiments in this area are occurring. Cross-domain FAIR is metadata-intensive, and generative AI may provide us with tools for meeting the increased demands for metadata, so long as we are able to ensure the accuracy of the information produced.

7.4. Packaging

There are situations where it is useful to assemble collections of related resources together for a specific purpose, and some form of packaging is necessary. For the purposes of archiving, we may need to assemble packages that are complete unto themselves, without reference to external links on the Web, etc. For dissemination purposes, it may be helpful to create a package for download or other technical purposes, combining relevant data and metadata.

[175] https://research.google/blog/croissant-a-metadata-format-for-ml-ready-datasets/
[176] https://www.bpmn.org/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

One possibility for performing this function which is used within the FAIR community is the Research Object Crates specification (RO-Crate).[177] RO-Crate has also been proposed as one possible reference implementation of the FAIR Digital Object Framework (FDOF), as presented at the FAIR Digital Objects Forum 2024.[178] RO-Crate can be employed for different, specific purposes, and there are several proposed uses of RO-Crate to fulfil specific functions, such as the Workflow TO-Create Profile (RO-Crate with CWL to describe provenance, etc.).[179]

CDIF has yet to explore the range of use cases for which this type of packaging functionality might be useful, and the requirements in terms of interoperability. From a practical perspective however, we understand that such requirements may well present themselves. This topic remains to be further explored.

7.5. Additional data formats

The current recommendations cover the case where the data are assumed to be in a tabular text format when being described in a granular, "integration-ready" fashion. While this is a common case for data being downloaded, it is by no means the only one.

Large amounts of data are available in other forms, whether this is in XML serialisations like SDMX-ML[180], different RDF vocabularies (DataCube[181], CSV on the Web[182], etc.), formats like NetCDF[183], HDF5[184], or Parquet[185], or one of many application-specific formats. Many of these formats also have embedded or associated expressions of metadata regarding the data, with various degrees of standardisation. Different data formats exist to support different use cases: some data sets are very large, and a compact format is needed; others are intended to be queried as parts of a larger graph, etc. In many cases, the different formats require dedicated software tools or libraries for processing.

CDIF recognises the need to be able to describe these data for FAIR use. There is currently no established practice for providing standard, cross-domain FAIR metadata for them, however. In some cases (as for NetCDF) there are proposed formats for exposing the metadata or embedding metadata in data files; there are JSON conventions for associating metadata with CSV on the Web, etc. What is needed is an agreed convention for describing the metadata (as CDIF already proposes) but with some indication as to how the data themselves are formatted and can be processed.

[177] https://www.researchobject.org/ro-crate/
[178] https://research.manchester.ac.uk/en/publications/ro-crates-as-a-practical-implementation-of-fair-digital-object-to
[179] https://w3id.org/workflowhub/workflow-ro-crate/
[180] https://github.com/sdmx-twg/sdmx-ml
[181] https://www.w3.org/TR/vocab-data-cube/
[182] https://csvw.org/
[183] https://www.unidata.ucar.edu/software/netcdf/
[184] https://www.hdfgroup.org/solutions/hdf5/
[185] https://parquet.apache.org/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Consolidated references and bibliography

Bolton, S. (2022). CESSDA Data Access Policy (Version 2). Zenodo. https://doi.org/10.5281/zenodo.6722000

Bonino, L., Guizzardi, G., Sales, T., (2022) 'FAIR Digital Object Framework Documentation', https://fairdigitalobjectframework.org/

Cox SJD, Gonzalez-Beltran AN, Magagna B, Marinescu M-C (2021) Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17(6): e1009041. https://doi.org/10.1371/journal.pcbi.1009041

Dam, T., Krimbacher, Al, Neumaier, S. (2023). Policy Patterns for Usage Control in Data Spaces. Sem4Tra: Fifth International Workshop On A Semantic Data Space For Transport, 20 September, 2023, Leipzig, Germany. https://arxiv.org/pdf/2309.11289

Gregory A, Wackerow J, Orten H (2023) Reuse and Reproducibility: Describing Cross-Domain Research Data in the Science Project Climate Neutral and Smart Cities. ARPHA Preprints. https://doi.org/10.3897/arphapreprints.e115047

Gregory, A., Hodson, S., Wackerow, J. (2021). The Role of DDI-CDI in EOSC: Possible Uses and Applications. Zenodo. https://doi.org/10.5281/zenodo.4707263.

Esteves, Beatriz, and Víctor Rodríguez-Doncel. "Analysis of ontologies and policy languages to represent information flows in GDPR." Semantic Web Preprint (2022): 1-35 https://www.researchgate.net/publication/361160523_Analysis_of_ontologies_and_policy_languages_to_represent_information_flows_in_GDPR

Esteves, B., Rodríguez-Doncel, V., Pandit, H. J., Lewis, D. (2023). Semantics for implementing data reuse and altruism under EU's Data Governance Act. In: 19th International Conference on Semantic Systems, 20-22 Sep 2023, Leipzig, Germany. http://dx.doi.org/10.3233/SSW230015

European Commission, Directorate-General for Research and Innovation, Cost-benefit analysis for FAIR research data – Cost of not having FAIR research data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/02999

European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality – Final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524

Mingfang Wu, Stephen M. Richard, Chantelle Verhey, Leyla Jael Castro, Baptiste Cecconi, Nick Juty; An Analysis of Crosswalks from Research Data Schemas to Schema.org. Data Intelligence 2023; 5 (1): 100–121. doi: https://doi.org/10.1162/dint_a_00186

Open Geospatial Consortium Abstract Specification - TOPIC 25 - ABSTRACT CONCEPTUAL MODEL FOR TIME - OGC doc 23-049 provides a conceptual framework for these different representations - (to be published) - draft https://portal.ogc.org/files/?artifact_id=107087&version=2

Pandit, Harshvardhan J., and Beatriz Esteves. "Enhancing Data Use Ontology (DUO) for Health-Data Sharing by Extending it with ODRL and DPV." Preprint on webpage at https://www. semantic-web-journal. net/system/files/swj3127. pdf (2023).

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Policy Patterns for Usage Control in Data Spaces 20 Sep 2023, https://arxiv.org/pdf/2309.11289.pdf

Rambla, J., Beltran, S., & D'Altri, T. (2023). European Genomic Data Infrastructure project (GDI) D8.4 Report on federated data access scenarios. Zenodo. https://doi.org/10.5281/zenodo.8208439

Weigel, T., Plale, B., Parsons, M., Zhou, G., Luo, Y., Schwardmann, U., Quick, R., Hellström, M., Kurakawa, K. (2018). RDA Recommendation on PID Kernel Information (Version 1). DOI: https://doi.org/10.15497/RDA00031

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

Wittner, R., Geiger, J., Müller, H., Frexia, F., Mascia, C., Exter, K., Cancio, I., & Holub, P. (2023). EOSC-Life Common provenance model for processing biological material, data generation and computational workflows. Zenodo. https://doi.org/10.5281/zenodo.8279525

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Appendix 1. Serialisation of CDIF metadata

JSON-LD has been chosen as the recommended serialisation format for CDIF metadata following our principle to use existing mainstream technology. The JSON format is widely used for data serialisation and popular with developers. JSON-LD adds additional syntax for the representation of linked data, compatible with existing JSON implementations so that integration with existing applications is relatively frictionless. Many metadata providers are using the schema.org[186] vocabulary with JSON-LD serialisation for metadata publication and interchange. Use of this format provides a low barrier to entry for data providers.

The JSON syntax is defined by the ECMA JSON specification[187], and JSON-LD is specified in the JSON-LD 1.1 recommendation[188] from the World Wide Web Consortium (W3C). This serialisation is designed for linked data applications that will translate the JSON into a set of {subject, predicate, object} triples that can be loaded into an RDF database for processing. The JSON-LD context binds JSON keys to URIs for more precise semantics, and the use of URIs to identify entities and property values in the metadata will maximise the linkage with resources on the wider Web to build an ever-expanding global knowledge graph.

A metadata record has two parts; one part is about the metadata record itself, the other part is the content about the resource that the metadata documents. The part about the record specifies the identifier for the metadata record, agents with responsibility for the record, when it was last updated, what specification or profiles the metadata serialisation conforms to, and other optional properties of the metadata that are deemed useful. The metadata about the resource has properties about the resource like title, description, responsible parties, spatial or temporal extent (as outlined in the Metadata Content Requirements section).

Schema.org includes several properties that can be used to embed information about the metadata record in the resource metadata: sdDatePublished, sdLicense, sdPublisher, but lacks a way to provide an identifier for the metadata record distinct from the resource it describes, to specify other agents responsible for the metadata except the publisher, or to assert specification or profile conformance for the metadata record itself.

There are two patterns that could be used to structure the two parts of the metadata record:

Option 1. The root object is the described resource:

```
{   "@context": "https://schema.org",
    "@id": "ex:URIforDescribedResource",
    "@type": "ImageObject",
    "name": "Picture of analytical setup",
    "description": "Description of the resource",
    "subjectOf": {
        "@id": "ex:URIforTheMetadata",
        "@type": "DigitalDocument",
        "dateModified": "2017-05-23",
```

[186] https://schema.org/
[187] https://www.ecma-international.org/publications-and-standards/standards/ecma-404/
[188] https://www.w3.org/TR/json-ld11/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
        "encoding": {
            "@type": "MediaObject",
           "dcterms:conformsTo": {"@id":"ex:cdif-metadataSpec"}
          },
        "about":{"@id":"ex:URIforDescribedResource"}
    }  }
```

Option 2: root object is the metadata record

```
{   "@context": "https://schema.org",
    "@id": "ex:URIforTheMetadata",
    "@type": "DigitalDocument",
    "dateModified": "2017-05-23",
    "encoding": {
          "@type": "MediaObject",
         "dcterms:conformsTo": {"@id":"ex:cdif-metadataSpec"}
          },
    "about": {
         "@id": "ex:URIforDescribedResource",
         "@type": "ImageObject",
         "name": "Picture of analytical setup",
         "description": "Description of the resource",
         "subjectOf":{"@id":"ex:URIforTheMetadata"}
       }   }
```

The rdf triples generated by these two approaches are identical, so if the metadata are always harvested to a triple store it makes no difference. However, allowing either approach would create interoperability problems for harvesters that are parsing the metadata as JSON-- the paths to the same metadata elements are different in the two approaches. It is our judgement that option two above (root object is the metadata) is more consistent with knowledge graph construction, CDIF thus recommends a JSON-LD serialisation in which the root object is the metadata record. If this is a problem for processing JSON-LD metadata as JSON, JSON-LD framing[189] can be used to generate a desired, consistent JSON tree structure for processing.

The recommended basic structure of the JSON-LD file is like this (using the 'root object is the metadata' approach). This pattern is used in the examples below:

```
{
    "@context": [
        "https://schema.org",
        {"dcterms": "http://purl.org/dc/terms/",
         "ex":"https://example.com/99152/"
        }
    ],
    "@id": "ex:URIforThisMetadataRecord",
    "@type": "DigitalDocument",
    "dateModified": "2017-02-03",
```

[189] https://www.w3.org/TR/json-ld11-framing/

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
    "encoding": {
            "@type": "MediaObject",
           "dcterms:conformsTo": {"@id":"ex:cdif-metadataSpec"}
         },
    "about": {
        "@id": "ex:URIforDescribedResource",
         "@type": ```{the type of the described resource}```,
         "dateModified": "2014-02-23"
..... other metadata content omitted     }
```

JSON keys prefixed with '@' are keywords defined in the JSON-LD specification[190] (see table below)

Keyword	Description
@context	The value of the context is an object that specifies a set of rules for interpreting the JSON-LD document. The rules can be specified inline in, or via a URI that identifies a context object containing a set of rules.
@id	A string that identifies the subject of the assertions in the JSON object that contains the @id key.
@type	An identifier for the definition of the structure of the JSON object that contains the @type key. The type determines what keys or values should be expected in the JSON object that contains the key. Values are types defined in the schema.org vocabulary. In the CDIF framework (and for compatibility with FDOF digitalObjectType), the schema:additionalType property should be used (see implementation table below)

In the example above, there is a 'dateModified' metadata assertion. It would translate into a triple like this:

```
ex:URIforThisMetadataRecord schema:dateModified "2017-02-03".
```

Which states that the metadata was modified (most recently) on 2017-02-23.

On the other hand, in the 'about' object, there is a statement:

```
ex:URIforDescribedResource schema:dateModified "2010-02-03".
```

Which states that the Described Resource was modified (most recently) on 2010-02-03. The distinct identifier for the metadata record allows statements to be made about the metadata separately from statements about the resource it describes. Note that the @type for the metadata node (root node) is 'DigitalDocument'. This is a schema.org type that corresponds broadly to the concept of Digital Object as used by the Fair Digital Object (FDO) community[191], recognizing that the metadata record is a digital object.

[190] https://www.w3.org/TR/json-ld11/#keywords
[191] Bonino, Guizzardi, Sales 2022 (https://fairdigitalobjectframework.org/)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

A1.1. Implementation of metadata content items

The following table maps the metadata content items described in the Metadata Content Requirements section to the schema.org JSON-LD keys to use in metadata serialisation. Example metadata documents follow. The 'Obl.' column specifies the cardinality obligation for the property; '1' means one value required; 1? means one value required, but can be a nil value; 1..* means at least one value is required; 0..* means the property is optional and more than one value can be provided. Properties implemented with a path that starts with /"about" are describing the resource, properties with path from / describe the metadata.

CDIF content item	Obl.	Schema.org implementation	Scope note
Metadata identifier	1	/"@id":{URI}	The URI for the metadata record should be the @id value for the 'subjectOf' element in the JSON instance document tree
Metadata profile identifier	1	"dcterms:conformsTo": {identifier}	Use Dublin Core terms property. The value for Base CDIF metadata is '**CDIF_basic_1.0**'. Different profiles extending this must define unique identifier strings to use here.
Metadata contact	1	/ "maintainer":{Person or Organization}	Should include a name and contact point (institutional email is best) for the agent responsible for metadata content. This is the contact point to report problems with metadata content. Person and Organization are Agent objects with various properties.
Metadata Date	1	/"dateModified" : {date time}	Date of most recent update to metadata content. Use ISO 8601 format.
Resource Identifier	1	/"about"/"@id":{URI} Optional /"about"/"identifier":{URI}	The URI for the resource described by the metadata record; @id value for the “about” element in the JSON instance document tree. Can include a separate /"about"/"identifier" element, but recommend just providing the URI.
Title	1	/"about"/"name":{string}	A set of words that should uniquely identify the described resource for human use, in the scope of the metadata catalogue containing this metadata record.
Distribution	1	/"about"/"url":{URL}	If metadata is about a single digital object. Recommend including /"about"/"encodingFormat" to indicate the format of the distributed Digital Object.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

CDIF content item	Obl.	Schema.org implementation	Scope note
	1..*	/"about"/"distribution": { "@type": "DataDownload", "contentURL": {URL },... }	If the metadata is about an abstract, non-digital, or physical resource that has multiple distributions, with different URL, encodingFormat, conformsTo properties. Each distribution is considered a distinct digital object. contentURL is required.
Distribution Agent	1?	/"about"/"contributor": {"@type": "Role", "roleName": "provider", "contributor":{Person or Organization}..}	If a simple digital object with a single download URL, or a resource with multiple distributions all from the same provider, contact point for the provider of the distribution.
	1?..*	/"about"/"distribution": [{ "@type": "DataDownload","provider":{Person or Organization} }...]	If there are multiple distributions with different providers, each distribution can have a separate provider
Rights	1..*	/"about"//"license":{text or URI} Or /"about"//"conditionsOfAccess":{text or URI}	URL to licence document or text explanation of restrictions on use. There might be multiple links to documents specifying related security, privacy, usage, sharing, etc... concerns.
Resource type	1	/"about"/"@type":{schema.org type}	If the Schema.org resource types[192] are specific enough to scope the metadata record, use those.
	0..*	/"about"/"additionalType": [{DefinedTerm or URI}, ...]	If a more specific resource type needs to be specified, add a text or URI value here that identifies the type. MUST be consistent with the @type. To simplify parsing, always encode as an array.
Description	1?	/"about"/"description": {string}	Free text, with as much detail as is feasible
Originators	1?..*	/"about"/"creator" : [{Person or Organization}, ...]	The value is a schema.org person or organisation. To simplify parsing, always encode as an array. Use ORCID or other PID to describe person or organisation where possible
Publication Date	0..1	/"about"/"datePublished" : {date time}	Date on which the resource was made publicly accessible. Use ISO 8601 format.
Modification Date	1?	/"about"/"dateModified" : {date time}	Date of most recent update to resource content. If Publication date is not provided, defaults to the Modification Date. Use ISO 8601 format.

[192] https://schema.org/docs/full.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

CDIF content item	Obl.	Schema.org implementation	Scope note
GeographicExtent (named place)	0..*	/"about"/"spatialCoverage": { "@type": "Place", "name": {string} or {schema:DefinedTerm} }	To specify location with place names; if the names are from a gazetteer, use the schema:DefinedTerm to provide a name, identifier, and inDefinedTermSet to fully document the concept.
GeographicExtent (bounding box)	1?	/"about"/"spatialCoverage": { "@type": "Place", "geo": { "@type": "GeoShape", "box": "39.3280 120.1633 40.445 123.7878" } } }	For bounding box specification of the spatial extent of resource content. See ESIP SOSO for details[193]. Recommend including only one bounding box; behaviour of harvesting clients when multiple geometries are specified is unpredictable.
GeographicExtent (point location)	0..1	/"about"/"spatialCoverage": { "@type": "Place", "geo": { "@type": "GeoCoordinates", "latitude": 39.3280 "longitude": 120.1633 } } }	For a point location specification of the spatial extent of resource content. Recommend including only one point; behaviour of harvesting clients when multiple geometries are specified is unpredictable.
GeographicExtent (other serialisation)	0..*	/"about"/"geosparql:hasGeometry": { "@type": "sf#Point", "geosparql:asWKT": "@type":#wktLiteral", "@value":"POINT(-76 -18)"}, "Geosparql:crs": {"@id":"CRS84"} }	Optional geographic extent using other more interoperable geometries, e.g., GeoSPARQL, see Ocean InfoHub[194]. (Note URIs in example are truncated...) Other geometry schemes might be specified in a specific domain profile, e.g. for atmospheric, subsurface data, or local coordinate systems.
Variable measured	1?..*	/"about"/"variableMeasured": [{ "@type":"PropertyValue", "@id": "astm:var0011", "propertyID": ["pato:PATO_0000025","astm:prop/0405"], "name": "hostMineral", "description": "...." }....]	Follow ESIPfed Science on Schema.org recommendation[195], see also discussion for representing more complex data structures in ESIPfed Experimental[196]. Variable must have a name and description, should have a propertyID with URI for the represented concept. The URI in the propertyID provides the semantic linkage for meaning of the variable. Applies to datasets, for other resource value is 'nil:notapplicable'.
Keyword	0..*	/"about"/"keywords": [{string}, {"@type":"DefinedTerm", "name": "OCEANS", "inDefinedTermSet": "gcmd:sciencekeywords", "identifier": "gcmd:concept/916b....6167d" },....]	Implement with text for tags, and schema:DefinedTerm for keywords from a controlled vocabulary. The DefinedTerm approach is used to represent concepts.

[193] https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Dataset.md#bounding-boxes
[194] https://book.oceaninfohub.org/thematics/spatial/README.html#simple-geosparql-wkt
[195] https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Dataset.md#variables
[196] https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Experimental.md#AdvancedVariableValueType

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

CDIF content item	Obl.	Schema.org implementation	Scope note
Temporal coverage	1?	/"about"/"temporalCoverage": "2018-01-22"	Calendar data or clock time instant use ISO8601 encoding
		/"about"/"temporalCoverage": "2012-09-20/2016-01-22"	Calendar data or clock time interval use ISO8601 encoding
		/"about"/"temporalCoverage": [{ "@type":"time:ProperInterval", "time:intervalStartedBy": "isc:LowerDevonian, "time:intervalFinishedBy": "isc:LowerPermian" }]	Time ordinal era interval, use owl:time namespace, time: http://www.w3.org/2006/time#. This example uses International chronostratigraphic chart, isc. See https://perio.do/en/ for identifiers for many other named time intervals.
Other related agents - simple contributor	0..*	/"about"/"contributor": [{Person or Organization}, ...]	Recognition for others who have contributed to the production of the resource but are not recognized as authors/creators.
related agent with role		/"about"/"contributor": {"@type": "Role", "roleName": "Principal Investigator", "contributor": {"@type": "Person", "@id": "https://orcid.org/...", "name": "John Doe", "affiliation": {"@type": "Organization", "@id": "https://ror.org/...", "name": "..." }, "contactPoint": {"@type": "ContactPoint", "email": "john.chodacki@ucop.edu" } } }	To assign roles to contributors like editor, maintainer, publisher, point of contact, copyright holder (e.g. DataCite contributor types), use the rather convoluted role construction defined by schema.org
Related resources	0..*	/"about"/"relatedLink": [{"@type":"LinkRole", "linkRelationship": "...", "target: {"@type": "EntryPoint", "encodingType": "text/html", "name": "...", "url": "https://example.org/data/stations" } }]	Use schema.org relatedLink with a LinkRole value, and the link URL in a 'target' EntryPoint object. These properties expect WebPage and Action as their domain, so the schema.org validator will throw a warning (not an error). Related resource links are useful for evaluation and use of data, but because of the wide variety of relationship possibilities, difficult to use in general search scenarios. Use a soft-type implementation, with a link relationship type using a schema:DefinedTerm, and a resolvable identifier for the relationship target.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

CDIF content item	Obl.	Schema.org implementation	Scope note
Funding	0..*	/"about"/"funding" : { "@id": "URI for grant", "@type": "MonetaryGrant", "identifier": "grant id", "name": "grant title", "funder": { "@id": "ror for org", "@type": "Organization", "name": "org name", "identifier": ["other identifiers"] } }	Use schema.org encoding and science on schema.org pattern[197] . Other organisation properties can be included in the funder/Organization .
Policies	0..*	"publishingPrinciples": [{"@type": "CreativeWork"}....]	FDOF digitalObjectMutability, RDA digitalObjectPolicy, FDOF PersistencyPolicy. Policies related to maintenance, update, expected time to live.
Checksum	0..1	/"about"/"spdx:checksum" or /"about"/"distribution": [{ "@type": "DataDownload", "spdx:checksum": {URL },.. }...]	A string value calculated from the content of the resource representation, used to test if content has been modified. No schema.org property, follow DCAT v3 adoption of Software Package Data Exchange (SPDX)[198] property; The spdx:checksum object has two properties[199]: algorithm and checksumValue. If the resource is a single Digital Object, use the first pattern, if there are multiple distributions, the checksum is a property of each distribution/DataDownload.
Version	0..1	/"about"/"version":"string version identifier"	A string value that identifies the version of the described resource

A1.2. Implementation Patterns

- DefinedTerm. {label, schemename, conceptURI, schemeURI}. This is a pattern used for property values that are concepts defined in a controlled vocabulary, ontology, or similar semantic artefact. Values have a label, which is a string that will be meaningful to a human user, a 'schemename', which is a label that similarly identifies the source semantic resource in which the concept is defined, the conceptURI is a globally unique, resolvable identifier for the concept value; schemeURI is a globally unique identifier for the semantic resource in which the concept is defined.

- Identifier. {identifier scheme, identifier string, resolvable identifier string}. This pattern is for identifiers that are useful to scope in the context of a scheme. The identifier scheme is associated with some

[197] https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Dataset.md#funding
[198] https://spdx.org/rdf/terms/
[199] https://spdx.org/rdf/spdx-terms-v2.1/classes/Checksum___-238837136.html

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

authority (e.g. the International ISBN Agency) that manages unique identifiers within their scope. If the identifier can be associated with a resolver to create a resolvable identifier string, typically an HTTP URL with a resolver host name (e.g. https://n2t.net/) to which the identifier is suffixed to obtain a representation of the thing identified.

- AgentObject. {name, agenttype, identifier, contactpoint, if the agent is a Person, include affiliation}. This pattern is for specifying an Agent in the prove sense: An agent is something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity. Agents can be persons, organisations, or software-defined actors. Agents have a name for human recognition, a type, an identifier, contactPoint and affiliation. Machine agent contact points should be the accessible human who operates the environment running the machine agent.

- DistributionObject {contentURL, encodingFormat, conformsTo, distributionAgent}. A pattern for specifying information necessary or useful for implementing machine access to a Digital Object that is or represents a resource of interest. Includes a URL for the Web location at which the Digital Object can be accessed, the format of the Digital Object, specifications or profiles to which the serialisation and content of the object conform, and the Agent responsible for the distribution platform. This agent is the contact point if there are problems accessing the distributed Digital Object.

- Nil values. For property values that are required but nilable, the logic is to use a nil value that indicates why the value is not provided if there is no value.

    - nil:missing. Value not provided in information used to create metadata, reason not provided
    - nil:unknown. The value is not known to the metadata creator.
    - nil:notapplicable. The property is not relevant to the resource being described
    - nil:withheld. The value is known, but is not public information and thus not provided.

A1.3. Examples

Example 1: simple digital object

This example shows a CDIF metadata record for a simple digital object-- a single image that is the resource of interest and the single representation of that resource:

```
{
    "@context": [
        "https://schema.org",
        {
            "dcterms": "http://purl.org/dc/terms/",
            "ex": "https://example.com/99152/"
        }
    ],
    "@id": "ex:URIforTheMetadata",
    "@type": "DigitalDocument",
    "dateModified": "2017-05-23",
    "provider": {
```

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
        "@type": "Organization",
        "name": "Joe's photo studio",
        "email": "metadata@joesphotostudio.org"
    },
    "dcterms:conformsTo": "ex:cdif-metadataSpec",
    "about": {
        "@id": "ex:URIforDescribedResource",
        "@type": "ImageObject",
        "name": "Picture of analytical setup",
        "description": "Description of what's in the picture, where and why taken.  note
that in this example, the described resource is  a Digital Object (not a a physical
object or abstract object with one or more representations implemented as Digital
Objects), so the resource metadata includes a URL that gets the object.",
        "license": "https://joesphotostudio.org/license.txt",
        "creator": "nil:unknown",
        "datePublished": "2012-07-18",
        "encodingFormat": "image/tif",
        "url": "https://repository.org/images/2423757.tif",
         "subjectOf":{"@id":"ex:URIforTheMetadata"}
    }
}
```

Example 2: a dataset with multiple distributions

The dataset is considered a non-digital resource-- it is a collection of data instances that can be represented in various ways. The metadata in this example must distinguish properties that are scoped to the dataset, independent of the representation (distribution), and properties that are distribution specific.

```
{
    "@context": [
        "https://schema.org",
        {"dct": "http://purl.org/dc/terms/",
         "spdx":"http://spdx.org/rdf/terms#"}
    ],
    "@id": "metadata:ark:/99152/URIforTheMetadataRecord",
    "@type": "DigitalDocument",
    "description": "NOTES on the metadata in the about section (below): conformsTo on
       the Dataset specifies the information model that underlies the data;  conformsTo
       in a DataDownload specifies the profile for the download-- access protocols,
       serialization scheme, vocabularies used, other conventions necessary to enable
       machine processing of the download",
    "maintainer": {
        "@type": "Person",
        "@id": "https://orcid.org/identifierForMetadataProducer",
        "name": "metadata creator-editor-steward name",
        "description": "person responsible for the metadata"
    },
    "dateModified": "2014-02-23",
    "encoding": {
```

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
        "@type": "MediaObject",
        "description": "this is about the encoding of the metadata in this metadata
record (a Digital Object...).",
        "contentSize": 687,
        "dcterms:conformsTo": "https://example.org/cdif-metadataSpec"
    },
    "about": {
        "@id": "ark:/99152/URIforTheDataset",
        "@type": "Dataset",
        "dcterms:conformsTo": "https://example.org/DatasetSpecification",
        "identifier": {
            "@type": "PropertyValue",
            "description": "this is redundant with @id, but makes identification of the
described dataset explicit",
            "propertyID": "https://registry.identifiers.org/registry/ark",
            "value": "ark:/99152/63v4yo3eeqepj0",
            "url": "https://n2t.net/ark:/99152/63v4yo3eeqepj0"
        },
        "publishingPrinciples": [
            {
                "@type": "CreativeWork",
                "url": "https://example.org/id/policy",
                "name": "Digital Object Policy",
                "description": "policies used in management of the described
                                resource..."
            }
        ],
        "distribution": [
            {
                "@type": "DataDownload",
                "contentURL": "https://example.org/datasets/1234567890.csv",
                "description": "A comma delimited text distribution of the data
                  following csv on the web W3C conventions (a Digital Object...). We do
                  not consider the URL that locates this particular Digital Object as
                  the identifier for the object, so this DataDownload is a blank node.
                  ",
                "spdx:checksum": {
                          "spdx:algorithm":"spdx:checksumAlgorithm_md5",
                          "spdx:checksumValue":"0BAA1B8"
                     },
                "contentSize": "12345 kb",
                "encodingFormat": "text/csv (base mime type)",
                "dcterms:conformsTo": [
                    "https://example.org/dataonthewebcsvprofile",
                    "https://fdof.org/fdofprofile"
                ]
            },
            {
                "@type": "DataDownload",
                "contentURL": "https://example.org/datasets/1234567890.rdb",
                "description": "A comma delimited text distribution of the data using
                  USGS RDB conventions (a Digital Object...). We do not consider the
```


'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
                  URL that locates this particular Digital Object as the identifier for
                  the object, so this DataDownload is a blank node. ",
               "spdx:checksum": {
                         "spdx:algorithm":"spdx:checksumAlgorithm_md5",
                         "spdx:checksumValue":"0F119B7"
                     },
               "contentSize": "11256 kb",
               "encodingFormat": "text/csv (base mime type)",
               "dcterms:conformsTo": [
                   "https://example.org/usgsRDBprofile",
                   "https://fdof.org/fdofprofile"
               ]
           }
       ],
       "dateModified": "2010-02-03",
       "subjectOf":{"@id":"metadata:ark:/99152/URIforTheMetadataRecord"}
   }
}
```

Example 3. Item list with a collection of metadata records:

This is an example file containing multiple records for harvesting, using the schema.org ItemList. The metadata record content is abbreviated.

```
{
    "@context": [
        "https://schema.org",
        {
            "dcterms": "http://purl.org/dc/terms/",
            "ex": "https://example.com/99152/"
        }
    ],
    "@type": ["ItemList"],
    "@id": "https://example.org/id/graph/X",
    "name": "Example CDIF Metadata collection",
    "description": "...",
    "itemListOrder": "https://schema.org/ItemListUnordered",
    "numberOfItems": 3,
    "itemListElement": [
        {
            "@id": "ID_for_this_metadata_record1",
            "@type": "DigitalDocument",
            "dcterms:conformsTo": "https://example.org/cdif-metadataSpec",
            "about": {
                "@id": "https://example.org/id/XYZ",
                "@type": "ImageObject",
               ....
            }
        },
        {
            "@id": "metadata:10.5878/tnzz-m331",
            "@type": "DigitalDocument",
            "dct:conformsTo": "https://example.org/cdif-metadataSpec",
```

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

```
            "about": {
                "@type": "Dataset",
                "@id": "https://doi.org/10.5878/tnzz-m331",
                ...
                "distribution": [
                    {
                        "@type": "DataDownload",
                        "name": "ID 006 - Vero, Beskattning av vind- ...",
                        "contentUrl": "https://snd.gu.se/en/catalogue/d...",
                        ...
                    },
                    {
                        "@type": "DataDownload",
                        "name": "ID 042 - Nye regler om stigning i ...",
                        "contentUrl": "https://snd.gu.se/en/catalogue/do...",
                        "encodingFormat": "application/pdf",
                    }
                ],
                "subjectOf": {"@id": "metadata:10.5878/tnzz-m331"}
            }
        },
        {
            "@id": "ex:URIforTheMetadata",
            "@type": "DigitalDocument",
            "dateModified": "2017-05-23",
            "provider": {
                "@type": "Organization",
                "name": "Joe's photo studio",
                "email": "metadata@joesphotostudio.org"
            },
            "dct:conformsTo": "ex:cdif-metadataSpec",
            "about": {
                "@id": "ex:URIforDescribedResource",
                "@type": "ImageObject",
                ...
                "subjectOf": {"@id": "ex:URIforTheMetadata"}
            }
        }
    ]
}
```

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Appendix 2. Mapping from CDIF metadata to RDA PID kernel attributes

The Fair Digital Object Framework (FDOF) specifies that a Digital Object identifier (PID) can be resolved to obtain a PID kernel record. For the content of the PID kernel record we are following the RDA Recommendation on PID Kernel Information (Weigel et al, 2018[200]). The implementation approach for supplying PID Kernel information records[201] associated with digital object identifiers is an architecture decision, not specified in this version of the CDIF framework. However, the information necessary to produce such metadata to implement the FDOF conventions is included in the CDIF metadata implementation, except for embedding of thumbnails or other data objects directly in the metadata digital object.

FDO field	CDIF schema.org	Scope Notes
FDO Creator	"creator" : [{Person or Organization}, ...]	Agent responsible for creating the FDO (and implicitly issuing the FDO PID) [NOTE-- creator of content and identifier registration are not necessarily the same agent]
FDO Responsible Organisation (Resource)	"provider":{Person or Organization}	note that this can be another organisation than the PID issuer. If the agent is an organisation, the value is taken from the ROR registry value domain (or other with namespace id). Implement as responsible part with an Agent (name, ID, contactInformation)
RDA checksum	"spdx:checksum":	Checksum of object contents. Checksum format is determined via the attribute type referenced in a Kernel Information record. Called etag in PubCom-PR-PIDProfileAttributes-2.0[202]. The algorithm for checksum calculation should be defined in the definition of the object type or described in the resource description in this metadata. spdx value is an object with 'algorithm' and 'checksumValue'.
FDOF digitalObject-Mutability	/"about"/"publishingPrinciples":	This attribute indicates whether the included bit-sequence is mutable or immutable, and policies for when new version is created when some bits are changed. Principles apply at the DigitalObject/Distribution level.
FDOF Persistency-Policy	/"about"/"publishingPrinciples":	this attribute indicates what the intention of its creator is with respect to its life-time/maintenance; the value domain is a vocabulary like: {UNKNOWN, NONE, Years , or EarthCube ELT } (note: seems only partly covered by RDA digitalObjectPolicy.
RDA digitalObject-Policy	/"about"/"publishingPrinciples":	Pointer to a policy object that documents changes to the object or its Kernel Information record, including object access and modification policies. A caller should be able to determine the expected future changes to the object from the policy, which are based on managed processes the object owner maintains.

[200] https://www.rd-alliance.org/system/files/RDA%20Recommendation%20on%20PID%20Kernel%20Information_final.pdf
[201] See sections 3.1 and 3.2 in https://fairdigitalobjectframework.org/
[202] https://docs.google.com/document/d/1QVUR6vlp6s6LxZndMslym9pmM90rFUR6__q3cuXSQf8/edit#heading=h.z337ya

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FDO field	CDIF schema.org	Scope Notes
FDOF Responsible-Organisation (Technical Management)	/"about"/"maintainer" : [{Person or Organization}, ...]	After creation, the same or another organisation will be responsible for further management of the FDO. The Responsible Organisation equals the FDO Creator if available by default (mandatory attribute) [equate with Resource point of contact]
FDOF rightsRecord	/"about"//"license":{text or URI}Or/"about"//"conditionsOfAccess":{text or URI}	This is a pointer to a resource that specifies access permissions. Include: FDOF licenceConditions: that links to one or more formal specifications about licences such as CC-x; FDOF transactionRecord: a resource that includes contractual information.
FDOF ScientificDomain	/"about"/"keywords":[{string} or schema:DefinedTerm]	indicator of the scientific domain the FDO refers to. This ensures compliance with the FAIR principles, which are per definition applicable at the domain level. This attribute is required since different mandatory attributes may be required at the domain-level.
Profile	/ "dcterms:conformsTo": {identifier}	The Kernel metadata profile is a schema that determines the attribute requirements for FDO metadata beyond the base requirements. In the FDO world, the kernel profile specifies Kernel information about the resource associated with an identifier. Schema.org does not have a 'conformTo' property so follow DCAT v3 using the Dublin Core Terms property.
FDOF digitalObjectType	/"about"/"@type":{schema.org type}	The kind of resource associated with an identifier. The type implies a schema that dictates the format, information model, and profile conventions for the resource representation contained in the identified digital object. Use appropriate Schema.org type for @type property, the additional type should be from a controlled vocabulary.
	/"about"/"additionalType":[{schema:DefinedTerm or URI}, ...]	Expect to use a CDIF recommended vocabulary here
RDA dateCreated	/"about"/"datePublished": {date time}	Date (and optional time) the Digital object was created
RDA dateModified	/"about"/"dateModified" : {date time}	If the DO bit sequence is mutable, specify the last date/time of object modification. Must be consistent with etag and current version number.
RDA version	/"about"/"version" : {string}	If tracked, a version for the object, which must follow a total order. Mandatory for all objects with at least one predecessor version.
RDA digitalObjectLocation	/"about"/"url":{URL}	If the FDO has a digital representation, it is mandatory that the PID record specifies the location where the FDO can be retrieved, either as an URL or a PID. This is URL in a metadata record for which the target resource is a digital object, or the contentURL or accessURL if the target resource is a non-digital object with one or more distribution representations. Since FDO PID identifies a

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

FDO field	CDIF schema.org	Scope Notes
		digital object, there is only one distribution, so use the simple schema:url.
FDOF operationInfo	Not implemented by CDIF v1.0	Some communities want to include a payload information such as a thumbnail image in the case of DiSSCo's Digital Enhanced Specimen FDO.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Appendix 3. Mapping from signposting relations to CDIF metadata elements

The signposting approach embeds links in HTML pages or in the HTTP header that is included in the response to HTTP GET or HEAD requests. These links provide some basic metadata information, as well as links to more complete metadata, possibly in various formats or profiles[203]. Given an identifier that can be dereferenced using a server that implements signposting, an agent can obtain some information about the resource(s) that can be accessed by dereferencing the identifier.

If a provider has generated CDIF metadata records, they have sufficient information to generate signposting links. If they can add HTML elements in landing pages, or HTTP elements in the server GET and HEAD responses, then implementing signposting is an easy win, providing value for some users. The table below shows the mapping from CDIF metadata elements to the targets for signposting links with each REL type.

Signposting Relation Type	CDIF schema.org element	Description
author	/"about"/"creator"/"@id"	The target of the link is a URI for an author of the resource that is the origin of the link.
cite-as	/"about"/"@id"	The target of the link is a persistent URI for the resource that is the origin of the link.
describedby	/"about"/"subjectOf"/"@id"	The target of the link provides metadata that describes the resource that is the origin of the link.
describes	/"about"/"@id"	The origin of the link is a resource that provides metadata that describes the resource that is the target of the link. It is the inverse of the describedby relation type.
type	/"about"/"@type": {schema.org type}; use "additionalType" for more specific typing	The target of the link is the URI for a class of resources to which the resource that is the origin of the link belongs.
license	/"about"/"license"/"@id"	The target of the link is the URI of a licence that applies to the resource that is the origin of the link.
item	/"about"/"relatedLink"/ /{"linkRelationship": "hasPart", "target": { "url":"http://someurl...", "contentType": "resource type of the item"}}	The origin of the link is a collection of resources, and the target of the link is a resource that belongs to that collection. It is the inverse of the collection relation type. In CDIF, contentType and encodingType properties can provide details about the kind of linked collection item.
collection	/"about"/"relatedLink"/ / {"linkRelationship": "IsPartOf", "target": { "url":"http:someURL.." "contentType": "Collection"}}	The origin of the link is a resource that belongs to a collection and the target of the link is the collection to which it belongs. It is the inverse of the item relation type. In CDIF, the target name and description (not in example to left) can provide additional information about collection.

[203] https://signposting.org/FAIR/#level1

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Appendix 4. Metadata for making data 'integration-ready': DDI-CDI classes and properties

This appendix lists the classes discussed in the text above, and indicates which properties must or may be used when these classes are instantiated.

It should be noted that while the Data Set and Data Structure classes appear in the DDI-CDI model, and are used in the diagrams above, these are never directly instantiated. Instead, their sub-classes will be used: Wide Data Set/Wide Data Structure, Long Data Set/Long Data Structure, and Dimensional Data Set/Dimensional Data Structure. For this reason, Data Set and Data Structure are not listed here.

The identifiers of each metadata declaration (or instance) are assumed to be the URI used in the JSON-LD. If other identifiers are to be associated with the instances, all of the classes listed can also use an identifier property as per the DDI-CDI specification. This is not required, however, as it is assumed that consumers of CDIF metadata will typically not be interested in non-public identifiers of this type.

Properties and associations are listed as "Required" or "Optional", followed by "Multiple" if more than one is allowed. Language-specific strings are allowed to be multiple, specific to language-locale pairs, using language maps as described in Example 71 of the JSON-LD specification[204].

All metadata instances will have @type referencing the DDI-CDI class, given in the "Type" field. Some instances will declare multiple types where appropriate. The classes are grouped by type, largely in the order they are mentioned in the steps above.

Note that associations and properties are distinct in DDI-CDI, because it is based on a UML formalisation where such a distinction is important (associations never link to literal values). They are listed separately here, for ease of reference to the DDI-CDI documentation.

Data Store

This represents a queryable service.

Type: cdi:DataStore

Properties:

name (Required) – a string assigning a name to the service. More complete description of the service can be provided using the fields in the CDIF Discovery recommendations.

Associations:

has (Required, Multiple) – links the Data Store to a Logical Record.

Logical Record

This describes the set of variables available in the service or data set at a logical level.

[204] https://w3c.github.io/json-ld-syntax/#string-internationalization

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Type: cdi:LogicalRecord

Properties:
None.

Associations:

- *organizes* (Optional, Multiple) – points to a Wide Data Set, Long Data Set, or Dimensional Data Set.
- *has* (Required, Multiple) – points to the Instance Variables making up the Logical Record.
- *isDefinedBy* (Optional, Multiple) – points to a Unit Type, Universe, or Population which provides a description of the subject of the Logical Record.

Wide Data Set

This represents a data set where each row in the table is a record, holding a set of values for variables, represented by columns.

Type: cdi:WideDataSet, cdi:PhysicalDataSet

Properties:

- *name* (Required) – a string value holding the name of the data set. May be provided in multiple, language-specific versions.
- *physicalFileName* (Required) - a URL providing the location of the data set.

Associations:

- *isStructuredBy* (Required) – points to an instance of cdi:WideDataStructure describing the structure of the data set.

Long Data Set

This represents a data set where each row in the table is one or more measurements, with another column for each measure in the row that indicates which logical variable is being measured. The variable being measured is specified for each row, and can vary from row to row.

Type: cdi:LongDataSet, cdi:PhysicalDataSet

Properties:

- *name* (Required) – a string value holding the name of the data set. May be provided in multiple, language-specific versions.
- *physicalFileName* (Required) - a URL providing the location of the data set.

Associations:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *isStructuredBy* (Required) – points to an instance of cdi:LongDataStructure describing the structure of the data set.

Dimensional Data Set

This represents a data set where each measurement and associated attributes are defined by a set of dimensions. The data set can be understood as a "cube". (See the DDI-CDI documentation for a more complete description.)

Type: cdi:DimensionalDataSet, cdi:PhysicalDataSet

Properties:

- *name* (Required) – a string value holding the name of the data set. May be provided in multiple, language-specific versions.
- *physicalFileName* (Required) - a URL providing the location of the data set.

Associations:

- *isStructuredBy* (Required) – points to an instance of cdi:DimensionalDataStructure describing the structure of the data set.

Wide Data Structure

This is a description of the structure of a Wide Data Set.

Type: cdi:WideDataStructure

Properties:

- None.

Associations:

- *has* (Required) – Points to the cdi:PrimaryKey used to identify records in the data set.
- *has* (Required, Multiple) – References the components (variables) which make up the records: links to cdi:IdentifierComponent, cdi:AttributeComponent, cdi:MeasureComponent.

Long Data Structure

This is a description of the structure of a Long Data Set.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Type: cdi:LongDataStructure

Properties:

- None.

Associations:

- *has* (Required) – Points to the cdi:PrimaryKey used to identify records in the data set.
- *has* (Required, Multiple) – References the components (variables) which make up the records: links to cdi:IdentifierComponent, cdi:AttributeComponent, cdi:MeasureComponent, cdi:DescriptorComponent, cdi:VariableValueComponent. Note that the components describe only the variables which are part of the presentation of the data set - other purely logical variables are referenced from the LogicalRecord associated with the Long Data Set and its components (see above).

Dimensional Data Structure

This is a description of the structure of a Dimensional Data Set.

Type: cdi:DimensionalDataStructure

Properties:

- None.

Associations:

- *has* (Required) – Points to the cdi:PrimaryKey used to identify records in the data set, made up of a set of cdi:DimensionComponents.
- *has* (Required, Multiple) – References the components (variables) which make up the records: links to cdi:DimensionComponent, cdi:AttributeComponent, cdi:MeasureComponent.
- *has* (Optional, Multiple) – Points to the cdi:ComponentPosition instances used to order the instances of cdi:DimensionComponent when the Primary Key is ordered.

Instance Variable

This represents a variable used in a data set. The full set of logical variables must always be described, so that every value in the data set is accounted for.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Type: cdi:InstanceVariable, cdi:RepresentedVariable [only if Represented Variables are not instantiated separately]

Properties:

- *name* (Required) – a string value holding the name of the variable. May be provided in multiple, language-specific versions.
- *definition* (Required) – a string value holding the definition of the variable. May be provided in multiple, language-specific versions.
- *describedUnitOfMeasure* (Optional) - a specification of the UOM, specified according to the "defined term" implementation pattern specified in Appendix 1.
- *unitOfMeasureKind* (Optional) - a specification of the quantity-kind for the UOM, specified according to the "defined term" implementation pattern specified in Appendix 1.

Associations:

- *measures* (Optional) – a link to an instance of cdi:Population.
- *takesSubstantiveValuesFrom (*Required) - a link to an instance of cdi:EnumeratedValueDomain, which in turn provides information about the typing of the recorded values for the variable.
- *takesSentinelValuesFrom (*Optional) - a link to an instance of cdi:SentinelvalueDomain, which in turn provides information about how flags, missing values, and similar notations are typed in the variable.

Represented Variable

This represents a reusable variable description.

Type: cdi:RepresentedVariable

Properties:

- *name* (Required) – a string value holding the name of the variable. May be provided in multiple, language-specific versions.
- *definition* (Required) – a string value holding the definition of the variable. May be provided in multiple, language-specific versions.
- *describedUnitOfMeasure* (Optional) - a specification of the UOM, specified according to the "defined term" implementation pattern specified in Appendix 1.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *unitOfMeasureKind* (Optional) - a specification of the quantity-kind for the UOM, specified according to the "defined term" implementation pattern specified in Appendix 1.

Associations:

- *measures* (Optional) – a link to an instance of cdi:Universe.
- *takesSubstantiveValuesFrom (*Required) - a link to an instance of cdi:SubstantiveValueDomain, which in turn provides information about the typing of the recorded values for the variable.
- *takesSentinelValuesFrom (*Optional) - a link to an instance of cdi:SentinelvalueDomain, which in turn provides information about how flags, missing values, and similar notations are typed in the variable.

Reference Variable

Reference Variables are presentational variables used only in the description of Long Data Sets. They contain values from the set of logical variables held in the data, and must have a representation which supports the full range of possible values from them.

Type: cdi:ReferenceVariable

Properties:

- *Name* (Required) – a string value holding the name of the variable. May be provided in multiple, language-specific versions.

Associations:

- *takesSubstantiveValuesFrom (*Required) - a link to an instance of cdi:SubstantiveValueDomain, which in turn provides information about the typing of the recorded values for the variable.
- *takesSentinelValuesFrom (*Optional) - a link to an instance of cdi:SentinelvalueDomain, which in turn provides information about how flags, missing values, and similar notations are typed in the variable.

Descriptor Variable

Descriptor Variables hold values which reference the logical variables in the data set, indicating which one the associated value in the corresponding Reference Variable is a measure/value for. Descriptor Variables are presentational variables found only in Long Data Sets.

Type: cdi:DescriptorVariable

Properties:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *Name* (Required) – a string value holding the name of the variable. May be provided in multiple, language-specific versions.

Associations:

- *takesSubstantiveValuesFrom (*Required) - a link to an instance of cdi:DescriptorValueDomain, which in turn provides information about logical variables to which each enumerated value corresponds, using the cdi:Descriptor class.

Substantive Value Domain

Substantive values reflect the definition of the variable which they populate, as opposed to being an indication of a process or system status ("missing," etc.). They may be enumerated or otherwise typed.

Type: cdi:SubstantiveValueDomain

Properties:

- None.

Associations:

- *recommendedDataType* (Required) - provides a link either to a skos:ConceptScheme (for enumerated values) or an XSD data type (see list in the text above). This describes the datatype of the values permitted as valid for populating the fields described by the variable.

Sentinel Value Domain

Sentinel values are those which indicate some process or system status or value which is not directly meaningful in terms of the variable definition of the field they populate. A typical example is flags for missing values.

Type: cdi:SentinelValueDomain

Properties:

- *None.*

Associations:

- *recommendedDataType* (Required) - provides a link either to a skos:ConceptScheme (for enumerated values) or an XSD data type (see list in the text above). This describes the datatype of the sentinel values permitted as valid for populating the fields described by the variable.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Descriptor Value Domain

This provides a list of instances of the Descriptor class, connecting the values found in a Long Data Set's Descriptor Variable with the logical variables for which values are provided in the Reference Variable.

Type: cdi:DescriptorValueDomain

Properties:

- None.

Associations:

- None.

Descriptor

An instance of the cdi: Descriptor class

Type: cdi:Descriptor

Properties:

- *content* (Required) - the value appearing in the Long Data Set, expressed as a string.

Associations:

- *identifies* (Required) – reference to the cdi:InstanceVariable which corresponds to the *content* property.

Unity Type

A Unit Type is the object or phenomenon serving as the subject of a data record, identified at a general level. If the values in a data set are "Age," "Sex", and "Occupation" the Unit Type would be "person". If the values are a measure of salinity and depth, the Unit Type might be "sea water". The value of an instance of cdi:UnitType is the general term for the unifying subject of a data record (an instance of cdi:LogicalRecord).

Type: cdi:UnitType

Properties:

- *name* (Required) – a string value holding the name of the variable. May be provided in multiple, language-specific versions.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *descriptiveText* (Optional) – a textual description of the Unit Type. May be provided in multiple, language-specific versions.

Associations:

- None.

Universe

An instance of cdi:Universe qualifies a Unit Type, to provide a more-specific description of the subject of the data, but excluding time and geography. If the Unit Type is “person,” the Universe might be “adults”. If the Unit Type is “sea water,” the Universe might be “sea water in temperate oceans”.

Type: cdi:Universe

Properties:

- *name* (Required) – a string value holding the name of the variable. May be provided in multiple, language-specific versions.
- *descriptiveText* (Optional) – a textual description of the Unit Type. May be provided in multiple, language-specific versions.

Associations:

- *uses* (Optional) – a reference to the instance of cdi:UnitType which the Universe qualifies.

Population

This is a fully qualified description of the subject of the data in a field or record. It extends the description provided by Unit Type and Universe by adding geographical and temporal bounds.

Type: cdi:Population

Properties:

- *name* (Required) – a string value holding the name of the variable. May be provided in multiple, language-specific versions.
- *descriptiveText* (Optional) – a textual description of the Unit Type. May be provided in multiple, language-specific versions.

Associations:

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *uses* (Optional, 2 Maximum) – a reference to the instance of cdi:Universe and/or cdi:UnitType which the Universe qualifies.

Identifier Component

The instance of cdi:IdentifierComponent is a reference to a variable which provides a unique identifier for the case being observed in the Logical Record. It is typically an assigned or official identifier for the unit, according to an external scheme (e.g., a taxpayer ID for a person, a sample identifier, etc.). The Identifier Component is a role attached to the variable to describe its function within the data structure - it is usually part of (and may be the entire) key for the record.

Type: cdi:IdentifierComponent

Properties:

- None.

Associations:

- *isDefinedBy* (Required) – A reference to an instance of cdi:InstanceVariable (or, if present, cdi:RepresentedVariable) which defines the Identifier Component variable.

Measure Component

The instance of cdi:MeasureComponent indicates that the referenced variable contains a measurement or observation value, or a derived value, within the Logical Record. Bote that cdi:MeasureComponent might be a simple value (string, number, term), or an object (e.g. an array of spectral values). For this profile, we are restricting discussion to simple values.

Type: cdi:MeasureComponent

Properties:

- None.

Associations:

- *isDefinedBy* (Required) – A reference to an instance of cdi:InstanceVariable (or, if present, cdi:RepresentedVariable) which defines the Measure Component variable.

Attribute Component

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Each instance of cdi"AttributeComponent indicates that the referenced variable holds a value which describes the values in the Measure Components or otherwise provides information which is not itself a measure/observation or derived value.

Type: cdi:AttributeComponent

Properties:

- None.

Associations:

- *isDefinedBy* (Required) – A reference to an instance of cdi:InstanceVariable (or, if present, cdi:RepresentedVariable) which defines the Attribute Component variable.

Dimension Component

The instances of cdi:DimensionComponent are references to the variables which are the components of a compound identifier, in which each variable is a single field - an axis - in a coordinate system addressing a location in a matrix. These variables are often categorical, but also commonly include time. Unlike other Components, they may be ordered using instances of cdi:ComponentPosition.

Type: cdi:DimensionComponent

Properties:

- None.

Associations:

- *isDefinedBy* (Required) – A reference to an instance of cdi:InstanceVariable (or, if present, cdi:RepresentedVariable) which defines the Attribute Component variable.

Component Position

Each instance of cdi:ComponentPosition indexes a Dimension Component within the Primary Key of a Dimensional Data Structure, by assigning an incrementing *value* to it (lowest number comes first in the sequence).

Type: cdi:ComponentPosition

Properties:

- *value* (Required) - An integer indicating the position of the Dimension Component within the sequence of the Primary key for the referencing Dimensional Data Structure.

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Associations:

- *indexes* (Required) – a reference to the instance of cdi:DimensionComponent which is being ordered. Assigns the *value* to it as a position within the Primary Key.

Primary Key

The instance of Primary Key defines the set of Components which, taken together, provide the identifying values for a single logical record within a data structure. The values of the Primary Key are used to reference the non-Primary Key fields within the Logical Record, by additionally specifying the Component/Instance Variable wanted. Such fields include the Measure Components and non-identifying Attribute Components.

Type: cdi:PrimaryKey

Properties:

- None.

Associations:

- *isComposedOf* (Required, Multiple) – References to the instances of cdi:PrimaryKeyComponent making up the Primary Key.

Primary Key Component

Instances of cdi:PrimaryKeyComponent reference the Components which make up the primary Key in a Data Structure.

Type: cdi:PrimaryKeyComponent

Properties:

- None.

Associations:

- *correspondsTo* (Required) – A reference to the instance of cdi:IdentifierComponent, cdi:AttributeComponent, cdi:DimensionComponent, or cdi:VariableDescriptorComponent used as a Primary Key Component.

Variable Value Component

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Each instance of a cdi:VariableValueComponent points to a Reference Variable, which holds the values of several logical variables combined into a single field within a Long Data Structure. This Component is part of a presentational structure, as opposed to a logical one.

Type: cdi:VariableValueComponent

Properties:

- None.

Associations:

- *isDefinedBy* (Required) – A reference to an instance of cdi:ReferenceVariable.

Variable Descriptor Component

Instances of cdi:VariableDescriptorComponent are fields in a Long Data Structure which indicate the logical variable with which the value of the Variable Value Component/Reference Variable are associated.

Type: cdi:VariableDescriptorComponent

Properties:

- None.

Associations:

- *refersTo* (Required) – A reference to an instance of cdi:DescriptorVariable.

Physical Segment Layout

Each instance of cdi:PhysicalSegmentLayout describes the physical expression of a record in a data file or other storage expressed as a textual format. The Logical Record encoded is referenced. There is only a single Physical Segment Layout used to describe data files - all records are assumed to have the same layout.

Type: cdi:PhysicalSegmentLayout

Properties:

- *arrayBase* (Optional) - This is a literal value of either “0” or “1” indicating the starting number for the sequence of variables in the record as indexed by instances of cdi:ValueMappingPosition for each variable.

- *commentPrefix* (Optional) - A string used to indicate that a line in the input is a comment.

‘Global cooperation on FAIR data policy and practice’ (WorldFAIR) has received funding from the European Union’s Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *delimiter* (Optional) - The character used to separate fields in a delimited file (for example, the comma in a CSV file),
- *encoding* (Required) - A string specifying the character encoding (typically UTF-8). Values are the same as those used for the HTML *charset* attribute.[205]
- *escapeCharacter* (Optional) - Provides the string which is used to escape the quote character.
- *hasHeader* (Optional) - A boolean value indicated as "true" or "false" (case sensitive) which indicates that the number of initial rows in *headerRowCount* should be skipped, as they do not contain data values.
- *headerRowCount* (Optional) - An integer which indicates now many initial rows to skip as containing header information. Only used if *hasHeader* has a value of "true".
- *isDelimited* (Optional) - A boolean value indicated as "true" or "false" (case sensitive) which says whether the file is delimited. If "true", then the *delimiter* property will provide the delimiter. If "true" then *isFixedWidth* must have a value of "false" if provided.
- *isFixedWidth* (Optional) - A boolean value indicated as "true" or "false" (case sensitive) which says whether the file has fixed-width fields. If "true" then *isDelimited* must have a value of "false" if provided. The starting and ending positions of each field are provided by the instances of cdi:ValueMapping associated with each variable in the record, in relation to the *arrayBase* property.
- *lineTerminator* (Optional) - The character sequence used to terminate rows. Defaults to the array [CRLF, LF] as per the Metadata Vocabulary for Tabular Data.[206]
- *quoteChatacter* (Optional) - The string that is used around escaped cells, or null, set by the quoteChar property of a dialect description. The default is ".". See W3C Recommendation "Model for Tabular Data and Metadata on the Web".[207]
- *skipInitialSpace* (Optional) - A boolean value expressed as "true" or "false" (case-sensitive) indicating that - if "true" - white space at the beginning of a row or following a delimiter should be ignored.
- *trim* (Optional) - This field takes the values "Both", "End", "Neither", and "Start" describing the places where white space can be trimmed from rows. Default is "Neither".

Associations:

- *Formats* (Required) - specifies an instance of cdi:LogicalRecord.

[205] See https://www.w3schools.com/html/html_charset.asp.
[206] https://www.w3.org/TR/tabular-metadata/
[207] https://www.w3.org/TR/tabular-data-model/#parsing

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *has* (Required, Multiple) – A reference to an instance of cdi:ValueMapping.
- *has* (Required, Multiple) - A reference to an instance of cdi:ValueMappingPosition, one per Value Mapping.

Value Mapping

Instances of cdi:ValueMapping are used to indicate the connection of variables with a Physical Segment layout.

Type: cdi:ValueMapping

Properties:

- *decimalPositions* (Optional) - An integer indicating the number of decimal positions when the decimal separator is implied.
- *defaultDecimalSeparator* (Optional) - A string providing the character used as a decimal separator.
- *length* (Optional) - An integer specifying the length of a fixed-width field.
- *maximumLength* (Optional) - An integer indicating the maximum length of the field value.
- *minimumLength* (Optional) - An integer expressing the maximum length of the field value

Associations:

- *formats* (Required) - specifies an instance of cdi:InstanceVariable, cdi:RepresentedVariable, cdi:ReferenceVariable, ot cdi:DescriptorVariable.

Value Mapping Position

Instance of cdi:ValueMappingPosition provides the sequence of the mapped variables within the Physical Segment Layout.

Type: cdi:ValueMappingPosition

Properties:

- *value* (Required) – An integer indicating the relative position of the mapped variable in the sequence of the Physical Segment Layout. Uses the *arrayBase* property of the Physical Record Segment as a starting point.

Associations:

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

- *indexes* (Required) – A reference to an instance of cdi:ValueMapping

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.

Appendix 5. CDIF Working Group and Advisory Group

CDIF Working Group

Arofan Gregory (CODATA), Chair
Darren Bell (UKDA)
Dan Brickley (Independent Expert)
Pier Luigi Buttigieg (AWI, ODIS)
Michelle Edwards (University of Guelph)
Doug Fils (Ocean Experts)
Luis Gerardo Gonzalez Morales (UN Stats)
Pascal Heus (Postman)
Simon Hodson (CODATA)
Chifundo Kanjala (UNICEF)
Yann Le Franc (eScience Factory)
Lauren Maxwell (University of Heidelberg)
Steve Richard (Independent Expert)
Flavio Rizzolo (Stats Can)
Peter Winstanley (Semantic Arts)

CDIF Advisory Group

Pierre-Antoine Champin (W3C)
Franck Cotton (INSEE, Making Sense)
Joachim Wackerow (Independent Expert)
Steven McEachern (ADA)
Natalie Meyers (University of Notre Dame)
Tom Baker (Independent Expert, Dublin Core Metadata Initiative)
Alejandra Gonzalez Beltran (UK Atomic Energy Agency, previously STFC)
Stian Soiland-Reyes (University of Manchester)
Simon Cox (Independent Expert, OGC, previously CSIRO)
Jay Greenfield (CODATA)
Erik Schultes (GO FAIR Foundation)
Luiz Bonino (University of Twente)
Christine Kirkpatrick (San Diego Supercomputing Centre)
Stuart Chalk (University of Florida, IUPAC)

'Global cooperation on FAIR data policy and practice' (WorldFAIR) has received funding from the European Union's Horizon Europe project call HORIZON-WIDERA-2021-ERA-01-01, grant agreement 101058393.