Project Title	Fostering FAIR Data Practices in Europe
Project Acronym	FAIRsFAIR
Grant Agreement No	831558
Instrument	H2020-INFRAEOSC-2018-4
Topic	INFRAEOSC-05-2018-2019 Support to the EOSC Governance
Start Date of Project	1st March 2019
Duration of Project	36 months
Project Website	www.fairsfair.eu

D2.4 2nd Report on FAIR requirements for persistence and interoperability

Work Package	WP2 - FAIR Practices: Semantics, Interoperability, and Services
Lead Author (Org)	Leah Riungu-Kalliosaari (CSC)
Contributing Author(s) (Org)	Rob Hooft (DTL), Sylvia Kuijpers (SURF), Jessica Parland-von Essen (CSC), Jonas Tana (CSC)
Due Date	31.08.2020
Date	26.08.2020
Version	1.0 DRAFT NOT YET APPROVED BY THE EUROPEAN COMMISSION
DOI	https://doi.org/10.5281/zenodo.40016311631

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)

Abstract

This document is the second iteration of three reports on the state of FAIR in the European scientific data ecosystem, by the FAIRsFAIR project. This report focuses on providing relevant current information about persistent identifiers, semantic interoperability and technical implementations of the FAIR data principles. The report advises researchers, data-stewards and service providers to co-create and co-develop solutions case-by-case, but with a strong endeavour towards a larger FAIR ecosystem, seeking sustainable and cost-effective solutions.

Versioning and contribution history

Version	Date	Authors	Notes
0.9	03.08.2020	All authors	Draft for internal review
1.0	26.08.2020	Rob Hooft (DTL), Sylvia Kuijpers (SURF), Jessica Parland-von Essen (CSC), Jonas Tana (CSC)	Content ready

Disclaimer

FAIRsFAIR has received funding from the European Commission's Horizon 2020 research and innovation programme under the Grant Agreement no. 831558. 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.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Abbreviations and Acronyms

API	Application Programming Interface
DOI	Digital Object Identifier (schema)
DNS	Domain Name Server
CWL	Common Workflow Language
DCMI	Dublin Core Metadata Items
DNS	Domain Name System
DOIP	Digital Object Interface Protocol
DTR	Data Type Registry
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
FDO	FAIR Data Object
FREYA	A H2020 project aiming to extend the infrastructure for PIDs. Continuation of THOR
GEDE	Group of European Data Experts in RDA
HTTP	Hypertext Transfer Protocol
IF	Interoperability Framework
JSON-LD	JavaScript Object Notation for Linked Data
LOD	Linked Open Data
ORCID	Open Researcher and Contributor IDentifier
OWL	Web Ontology Language
PID	Persistent Identifier
PID KI	Persistent Identifier Kernel Information (metadata type)
RAID	Research Activity Identifier (schema)
RDA	Research Data Alliance
RDF	Resource Description Framework
SCHL	Shapes Constraint Language
SKOS	Simple Knowledge Organisation System
SPARQL	SPARQL Protocol and RDF Query Language
TTL	Terse RDF Triple Language (Turtle)
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
XML	Extensible Markup Language

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

About this document

In our first report[1], we described the FAIR landscape, taking into account semantic interoperability and persistence of research data management solutions. The report highlighted the diversity of the field. After finalising the first report, we organised a webinar[2] to present the results and invited the community to provide feedback[3,4]. We also organised two workshops with the aim of gathering feedback through focused discussions.[5] The first workshop was reserved for people that had expressed interest in contributing further after responding to the survey leading up to our first report. The second workshop was open to everyone.

In the two workshops, we gathered the participants' views about 1) the meaning of FAIRness, interoperability and persistent identifiers (PIDs); 2) FAIR technical implementation aspects such as linked data, and the resource description framework (RDF); and 3) general impressions of the first report and suggestions for improvement.

To complement the feedback and input from the community, we conducted desk research with a focus on recent developments related to implementing FAIR, with a particular focus on PIDs and metadata.

A few points have proven to be hard when trying to understand and implement the FAIR principles. This document takes a deep dive into a number of these difficulties: how to promote FAIR with PIDs, semantic interoperability and metadata. Each topic is described in a few sections, each targeting a specific audience. This way, we aim to provide researchers, data stewards and service providers (where possible) with information that can help them answer the question: "How do I choose the right techniques for FAIR data?".

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Executive Summary

This report is the second deliverable by the FAIRsFAIR project on technical implementation of FAIR principles. The first deliverable was a landscaping effort - to a general audience - that reviewed and documented commonalities and possible gaps regarding semantic interoperability, and the use of metadata and persistent identifiers across infrastructures. This report builds on the previous work and explores current developments to increase awareness on what good FAIRness means and how it could be promoted in practice.

The FAIR data principles have varying implications for different stakeholders. Thus, our aim is to provide an explanatory guide to researchers, data stewards and - where possible - service providers on the use of PIDs, metadata and semantic interoperability. We are presenting the information in sections geared towards a specific target audience i.e researchers, data stewards & service providers, with a focus on highlighting the aspects most relevant to the particular stakeholder group.

In order to achieve wide penetration and the potentially significant benefits of FAIR data, it is important for the development and implementation of FAIR data principles to be driven by researcher needs. Our main conclusions are as follows:

1. A generic solution for achieving FAIRness does not exist. The solutions should be selected and decisions made on a case-by-case basis. The assessment of FAIR data solutions should always start from the user needs but always with respect to the user's larger research community.
2. Every effort to make something FAIR should balance the investments needed to implement each FAIR principle, and the expected benefits of FAIRness to the scientific community.
3. In order to achieve a FAIR data ecosystem with sustainable PIDs, metadata and semantic artefacts, researchers, data stewards and service providers should work together on technical solutions.
4. Achieving Interoperability for both humans and machines requires a large investment, but it has promising benefits. Technology can solve a lot of the interoperability problems at a technical level - but this does not solve misunderstandings at the semantic level - humans still need to communicate, agree on terms and vocabularies. It is important to take advantage of existing frameworks to build cohesion.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Table of contents

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

1. Introduction

As part of the EOSC projects' ecosystem, the FAIRsFAIR - Fostering Fair Data Practices in Europe - project aims to supply practical solutions for the use of the FAIR data principles throughout the research data lifecycle. The FAIRsFAIR project lays emphasis on fostering a FAIR data culture and the uptake of good practices in making data FAIR (Findable, Accessible, Interoperable and Reusable), but there are still many discussions to have on what the implementations of the FAIR data principles actually mean, for instance for stakeholders such as researchers and data stewards. It is important to not only look at data management practices but also find solutions that are resilient over time.

This report is the second in a series of three that focuses on relevant solutions for implementing FAIR principles in practice. The first document was a gathering exercise of information aimed at a general audience. We reviewed the implementation of semantic interoperability and persistent identifiers in projects and landmarks listed by the European Strategy Forum on Research Infrastructures (ESFRI[6]) and also covered a broader perspective including, for instance, much of the important work done in the Research Data Alliance (RDA).

The feedback on the first report, the outcomes of the two workshops organised within this WP, and our observations on current developments suggest that persistent identifiers (PIDs) and metadata are subjects that need more discussions and work. The growing amounts of data have created an enormous need for practical data management solutions that align with the FAIR principles. In reality, data is often hard to discover (find) and difficult to reuse (accessible, interoperable & reusable), hence causing harm to the quality and efficiency of research. A recent study by the EOSC FAIR in practice working group[7] identified many existing technical difficulties with implementing FAIR in relation to repositories, interoperability, metadata and financial issues. In this document, we focus on describing how to overcome problems related to interoperability with PIDs and metadata, as these are important building blocks of a FAIR ecosystem and framework.

This document is intended to provide relevant information to researchers, data stewards and service providers - in order to shed light on some of the FAIR adoption principles and explain the most common misunderstandings. The information gathered aims at helping the readers in selecting the right techniques for implementing FAIR and understanding some of the essential practicalities related to PIDs, semantic interoperability and metadata. In order to address the specific level and context of the reader, some of the sections explicitly indicate the target audience. In order to know what their customers will be expecting, there is value for data stewards in the sections targeted for researchers, and there is value for service providers in reading what is written for data stewards.

We close the document with some generic conclusions for our three target audiences, researchers, data stewards and service providers, and that can also be useful for policy makers.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2. Overcoming the technical difficulties when implementing FAIR

For a data resource to be considered a FAIR digital object, it needs to be accompanied by persistent identifier(s) (PIDs) and metadata rich enough to enable it to be unambiguously identified and understood, used and cited, following metadata standards and vocabularies adopted by the related research community. In addition, the data needs to be represented in common, and open, formats.[8] The FAIR principles give guidance on the structure of digital objects and how they relate to each other. The principles realise this through the help of some repeating elements: unique and persistent identifiers, metadata, open protocols and interoperability are a few that have clear technical implications.[9]

2.1. Enabling FAIR with PIDs

Persistent identifiers are used to identify entities within the FAIR Ecosystem. They ensure unambiguous identification, and enable linking and referring to these entities across the ecosystem in a resilient way. The PID is an integral part of the FAIR data object. There are some basic features that a PID should have:

- it should be **globally unique**, i.e. nobody else in the world should use the same string to refer to anything else. In practice this means that a PID has a controlled syntax and a governed namespace (generally consisting of a name space indicator (prefix) and a local identifier (suffix)) and be issued and managed by a clearly specified registration authority;
- it should be **resolvable**, i.e. provide a way for both machines and humans to access the digital object itself, the state information and/or landing page (in current practice this means the identifier can be translated to a fully defined URI, at any moment, without the requirement that it resolves to the same URL over time);
- it should be **persistent**, i.e. remain unique and resolvable with a persistent syntax. The object it represents should ideally also be persistent, but even if that last persistence is broken the PID should guarantee not to be reused for any other object in the future.[10,11]

Few objects are actually completely persistent in the long term due to inevitable societal and technological changes. This means that all PIDs need active curation forever once published, if the trustworthiness of the PID system is to be ensured. As a PID should resolve on the internet with a common open protocol, for a human user today, this equals a webpage, and for a machine it means using HTTP..[12–14] The other standard mentioned is ITU X.1255: Framework for discovery of identity management information.[15] This requirement can and should also be formulated in a technology agnostic way; the PID should be recognizable as a PID by its intended user, be it a human user or a machine..[12–14]

2.1.1. Enabling FAIR with PIDs: by researchers

The PID forum[16] provides valuable information on getting started with PIDs for researchers and other stakeholders. A serious problem for FAIR data is what is sometimes called link-rot[17]: the fact that resources on the World Wide Web are moved over time, thereby invalidating references from outside to such resources. To mitigate this problem, the FAIR principles call for the managed use of PIDs: Persistent Identifiers. Important properties of these identifiers are that they are guaranteed to continue existing (persistent), that they are globally only in use to identify a single resource

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

(unique), and that there is a well-defined way to locate the actual resource based on the identifier (resolvable).

For research data, PIDs can be used to identify and locate entire published datasets, but they can also be valuable, already during a research project, for internal references between subsets of data and metadata or other data objects.

Researchers are as end users usually dependent on what the services e.g. software, data management, curation services etc, offer regarding data processing, lifecycle management and publication.[18] Metadata formats, the quality of reference metadata, version management, securing integrity, structural metadata and PID minting and allocation are all products of the services used. It is therefore good to be engaged in planning also the use of persistent identifiers with your data stewards and presenting your use cases to the service providers. From the researcher's point of view, there are two different PID contexts and they can serve as different use cases. We will discuss two use cases below. In both of them consistent use of identifiers and PIDs is relevant and helpful to your work especially in the long run.

The first use case is the **visibility of your work and outputs**. When reporting on your work, to funders, and publishing outputs, a basic level of FAIRness and PID use is sufficient to enable findability, simple citation and output registration with core descriptive metadata. This is the context of what is usually called *research information* (sometimes referred to as current research information). The most common and useful PIDs for this are the research output DOI and the ORCID for the creator(s)/contributor(s). There are also other systems available to identify other kinds of entities to help further linking of information, such as organisations or protocols. Funders and employers might for instance require linking to some other contextual reference data like lists of grants, funders and affiliated organisations. This kind of information is becoming more important, but the actual data quality is depending on the functionalities each service provides. If the services used for dataset publication or reporting don't require PIDs or don't offer reference (meta)datasets or integration with PIDs for these kinds of things, it is difficult for the researcher to provide this information in an unambiguous way.

The other use case for PIDs is the **management of the research data itself**. Here the PIDs can have different functions: (a) creating deep FAIR research datasets as *research outputs*, where all individual data elements are machine accessible, see panel F in Figure 1, or (b) when managing and documenting the actual workflow and data and related information *during research* to ensure reproducibility of research results.

Assessing and planning the use of PIDs might be different depending on which aspect is in focus and is best done by the researcher together with the data steward who can present different solutions or ways to use PIDs. This also requires that you, together, decide on where focus is put and what is good enough regarding for instance versioning, reproducibility or machine actionability. This is better validated prior to the start of the research, and then checked at regular intervals during the research. Their sustainability should also be evaluated both during the active research period and after the project.

Regarding machine accessibility there are different options depending on how developed data management and the domain infrastructure is and what is feasible and best serves the purposes of

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

the research community. Usually there needs to be good semantic interoperability (see the section about semantic interoperability) in place before machine actionability can be put in place. Panel F in figure 1 shows a situation where the data is "Deep FAIR" and accessible for both machines and humans on a data element level. In these kinds of solutions the data elements are curated so that the integrity of individual objects, as well as creation of subsets, is ensured and sustainable. Your research community should agree upon the appropriate depth of FAIRness of the data. Keep in mind that applying restrictions on machine accessible data elements (D) requires more extensive interoperability both legally and technically than open access cases, because you also have to create a way for identifying the users, their roles and access rights in an interoperable and automatic way before this can be implemented.

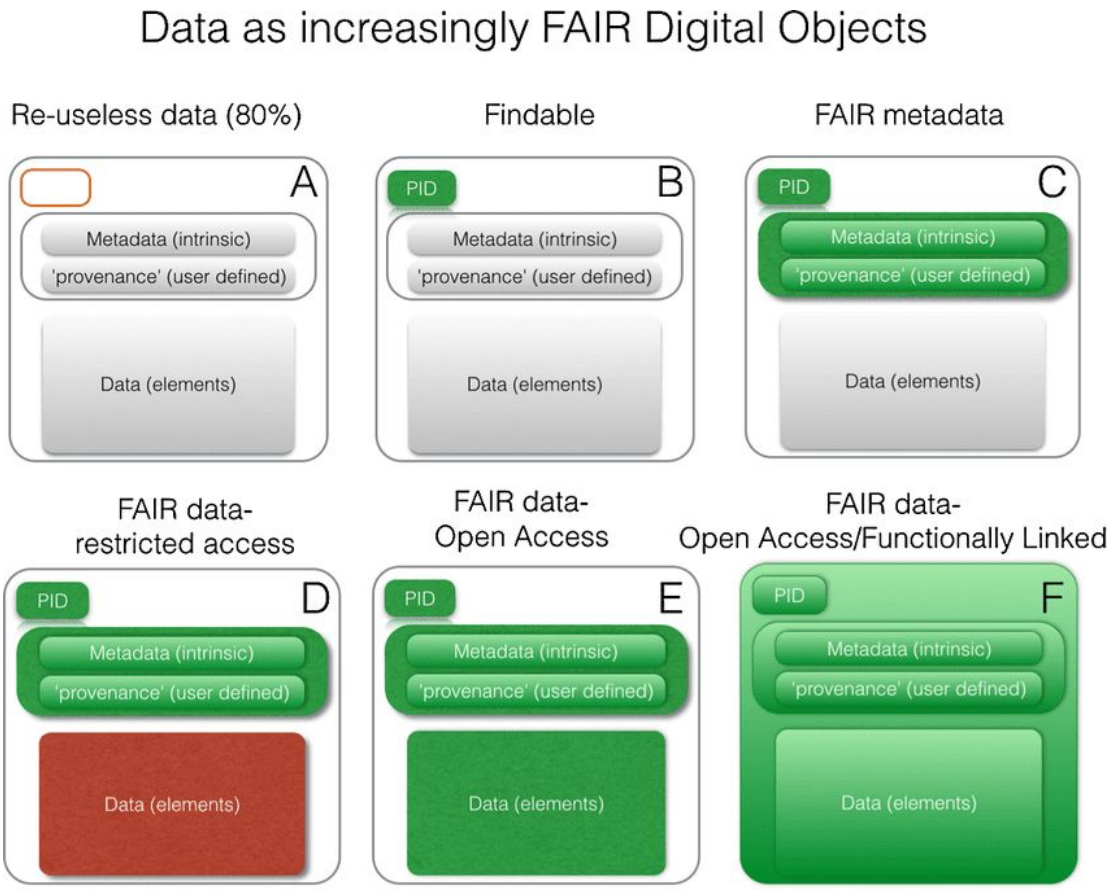

Figure 1 Different degrees of FAIR. Source: Mons et al 2017.[19] *We refer to "F" as "Deep FAIR".*

You can also use persistent and unique identifiers for supporting the creation of provenance metadata (provenance basically means 'how this came to be'), for instance by creating PIDs for sensors, instruments or workflows. Metadata and other requirements have to be defined by the research community when planning the information architecture and processes, but remember to look outside your own domain with the help of your data stewards.

When PIDs are used to identify digital objects during the research, they support workflows and automation in metadata creation and machine actionable metadata at earlier stages of the data lifecycle, for instance see this reference[20]. Metadata that is collected earlier and in an automated fashion like this is more likely to be complete and correct, and may save time, as compared to when

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

it needs to be recovered from scattered notes before data publication, which may be very time consuming and rushed due to nearing publication deadlines. These PIDs can act as anchor points in the data lifecycle. Code and software should also be considered - see further[21,22] and also FAIRsFAIR task 2.4 that is working with software as FAIR data objects. You can link or wrap PIDs, for instance, source data, software and outputs, to support the reproducibility of your research. PIDs can also be used in different ways when publishing actionable research outputs, thus supporting reproducibility. Documenting the research process and data provenance creates needs for identifying workflows. The Protocols service offers PIDs for protocols.[23] The Common Workflow Language (CWL) is a way to wrap the research process and it can also include manual activities.[24] There are different nascent ways of describing and structuring information about the processes and outputs of research, relevant ways among these are the Research Object Crate[25] and on a higher level the RAiD[26]. Today, it might often be the best or only solution to create a new FAIR data object (FDO) including all related PIDs to ensure sustainability, in which case the link to the source data might be vague. This in turn then reflects back to the research information level (see chapters 2.1.3 and 2.1.5).

It is NOT recommended that the researcher or any individual person is the PID owner, but this, as well as management, should be governed in a sustainable way.

As PIDs are a central element in creating FAIR data and ensuring registration of credits and scientific reproducibility, it is worthwhile spending some time and effort on exploring different solutions and options at an early stage of the research lifecycle and with the support of professional data stewards.

2.1.2. Enabling FAIR with PIDs: by data stewards

Choosing the right identifiers and PID schemas and systems needs good understanding of the aims and priorities of the research community. Granularity, proper levels of documentation and the need for reproducibility during the different stages of the research lifecycle should rather be discussed with the researchers sooner than later.
In this section we address a few specific types of PIDs in research that can broadly affect their use in a project. We attempt to help the data steward by identifying solutions for PIDs in general and encourage good practice so that data stewards can help assess their trustworthiness and whether they are aligned with the domain and project at hand. DataCite's PID Registry Service[27] provides an overview of different services related to PIDs. Also, there is an upcoming EOSC PID architecture guideline that will support implementation of the EOSC PID Policy. We highly recommend that data stewards familiarize themselves with these documents as it contains a lot of practical information that can help in choosing PID solutions. A tool that can be of some help is the PID service registry https://www.pidservices.org/.

2.1.2.1. PIDs for entire datasets vs for individual entities

For every dataset or data collection it is good to assess whether it should only have a high level PID (we refer to this as *Shallow FAIR*) such as for a whole database such as for a whole database, or if it is useful and stable enough for assigning PIDs for individual elements like data points or lines of that data too (*Deep FAIR*). A Deep FAIR dataset can have metadata describing separate elements, and would have every element findable and machine accessible. See also the section on assigning PIDS.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2.1.2.2. PIDs for data types

PIDs can help a dataset become more semantically interoperable (see also the section about semantic interoperability) through the use of a Data Type Registry (DTR). A DTR can be used to register for instance:

A. how variables in a dataset of the form w1, d2, temp, etc., correspond to real world notions of weight, distance, temperature, etc.
B. the measurement units associated with each of those dimensions, e.g., Kelvin, Celsius, or Fahrenheit in the case of temperature.
C. how those variables are grouped or packed together in datasets.[28]

The challenge in the use of PIDs in general, and Data Type Registries in particular, is that optimal interoperability requires that the same thing is addressed using the same PID in different contexts. Rather than defining new PIDs for an entity, be it a dataset, a piece of software, or a data type, you should always first check whether a PID already exists. Just creating new PIDs for every entity, without checking whether these entities already have existing PIDs, will not create interoperable data.

2.1.2.3. PIDs for instruments

The RDA Persistent Identification of Instruments working group (PIDINST) has collected use cases for persistent identification of instruments, and aims at aligning the collected metadata, and developing a metadata schema. In July 2019 the schema still contained a placeholder for the PID type as a suitable name for the instrument PID system still needs to be found.[29,30] Collaboration with DataCite resulted in a mapping of the PIDINST schema with the DataCite schema version 4.3. This now opens for using DataCite DOI for instruments.[31]

2.1.3. Creating sustainable interoperability with PIDs

Persistent identifiers have been discussed and developed within the scientific communities for decades, but there are still differences in how they are understood and implemented. An extensive effort to map this landscape was done in the GEDE consolidated assertions document in 2017.[11] Generally there seems to be quite a large consensus that governance, management, practice and cooperation are most important in creating good solutions for persistent identifiers, while the technical aspects are usually only solvable once agreement on terms and practice has been achieved.

Recently, the EOSC PID policy defined and discussed some of the important responsibilities and concepts regarding persistent identifiers.[10] The policy states that PIDs should be resolvable, i.e. machine actionable and that the PID metadata should contain only very limited information, called “kernel information”. As the existing services and systems are diverse, defining generic parts is by no means simple. The components listed in the EOSC PID policy are PID schemes and services, and the roles treated are authority, service providers, managers, owners and end users. An organisation might acquire one or several of the different roles (see figure 2).

Creating sustainable solutions for resolving the PIDs is an important part of a trustworthy and robust FAIR ecosystem. There are several different ways in which resolving can be done. The FREYA

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

project has described five different types of resolvers currently in use for different functions in the Internet and with different levels of usability for PID systems:

- Domain Name Service (DNS) resolver: Resolves a hostname to an IP address.
- Local resolver, e.g. load balancer, API gateway or web server: Redirects to a different host and/or path.
- Full resolver, e.g. handle system: Redirects to a URL either following a regular expression pattern, or a specific URL stored in the service.
- Meta-resolver, e.g. identifiers.org or n2t.net: Redirects to a URL following a regular expression pattern.
- Single-service resolver: some PIDs resolve to a single central resource, e.g. ORCID.[12]

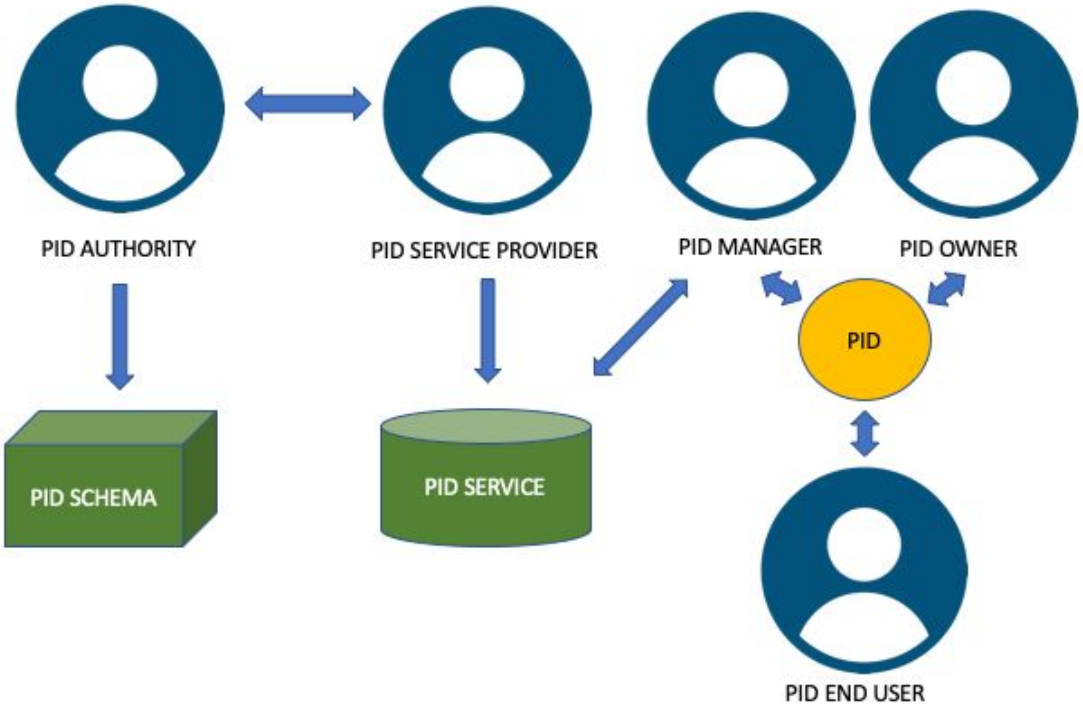

Figure 2. The PID owner is responsible for the integrity of the individual PID and can therefore even be an individual researcher. This might cause some practical problems with sustainability. Often the PID owner and manager are closely coupled.

The most sustainable of these are the full resolver and the meta-resolver as they provide the extra two-tier resolver layer that creates a buffer against organisational and technical changes and can offer robustness through networks. These have also been called first and second pattern identifiers.[32] The meta-resolver is useful when there are existing identifiers that for some reason are both sensible, sustainable and useful to integrate as (parts of) the suffix.

One important feature and goal for an efficient and resilient PID architecture, also pointed out in the EOSC Interoperability Framework, is that *any* resolver services should serve *all* PIDs (either directly or via redirection), much like the n2t.org service does today. There should also be clear guidelines on how to embed and nest different persistent identifier namespaces.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2.1.3.1. Constructing PIDs

A persistent identifier generally contains two parts (prefix and suffix), plus an initial namespace that makes it possible to recognize it as a PID and ensure uniqueness. The prefix is usually used by the PID authority and service to manage and control ownership of the PIDs when creating them. The suffix can then be either a hash or other string with or without semantics (the last case is also called *opaque*). In both cases, uniqueness has to be ensured when registering the PID, meaning that there is a guarantee that the same identifier has never been in use anywhere else for another goal.

Usually it is recommended that creating PIDs is done as early as possible in the data lifecycle. On the other hand, the occurrence of "zombie PIDs", referring to entities that never actually come into existence, can be a problem.[11] One way to manage the lifecycle could be to use existing non-global identifiers (e.g. those used inside a project) as (parts of) PID suffixes. This has to be done after carefully evaluating the risks posed by including semantics in the PID.

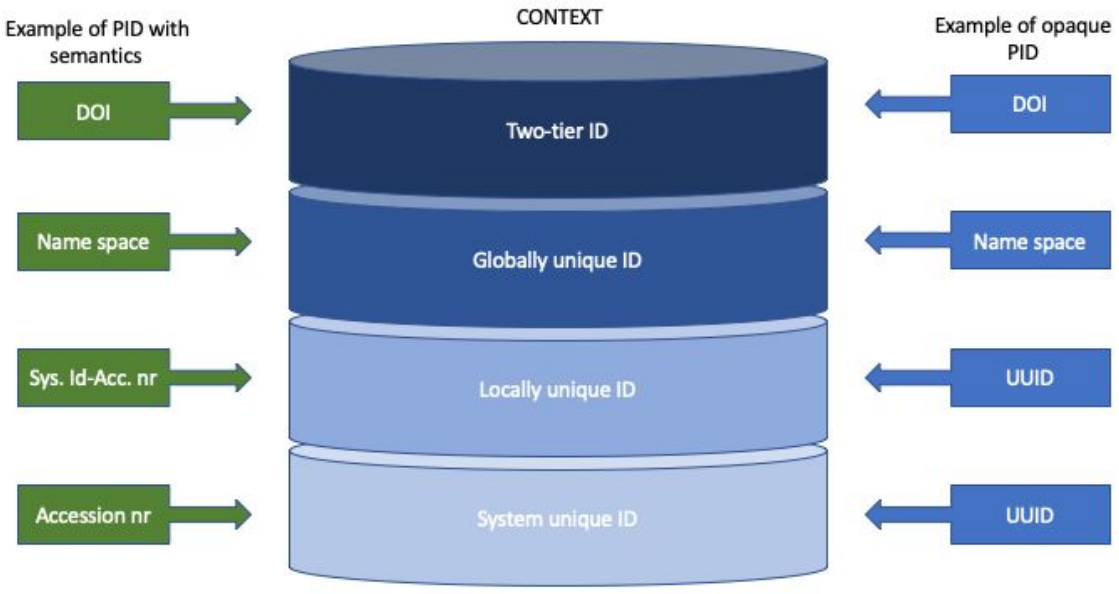

Figure 3. PIDs can be constructed in different ways. It is sometimes possible and useful to include elements of provenance or other IDs from a system or local context, but this should be done mindfully. The other option is to create a new completely opaque suffix. A namespace is always necessary, and the two-tier PID has a proxy layer that ensures machine actionability based on the namespace.

Often it is up to the owner to decide about the level of opaqueness of the PID. Maximal opaqueness is often recommended, in order to prevent the pressure to change the PID when components that are visible in the PID change for the object. An opaque PID also avoids its users to falsely assume that they can always be "constructed" following a recipe. Finally, an opaque PID hides the data type and provenance from the end user.[33] This principle optimizes interoperability by requiring every relation to be made explicitly. This is often represented with the hourglass analogy (see figure 5). A disadvantage of opaqueness is that complete opaqueness can make PIDs more difficult to identify for a human user, and impossible to interpret in case of complete failure of the PID system when the link to all related information is deleted.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Identifiers.org is one example of a service that builds PIDs with semantics and uses a meta-resolver. Their PIDs consist of an assigned unique prefix, followed by a colon and a provider-designated accession number (prefix:accession). The underlying Central Registry provides a centralized directory of these so-called Compact Identifiers. Resource maintainers can use a Prefix Registration Request form to request a prefix in Identifiers.org for their databases or services. Another example of a PID service, The California Digital Library's service Name to Thing (N2t) uses compact ID for several use cases while also promoting opaque PIDs generated by the Noid servers (Nice Opaque IDs) used by ARK.[34,35]

The data stewards play a very important role in supporting the researchers in finding the best PID solutions to enable FAIR data and scientific reproducibility. While the researchers can often describe their own needs and requirements, they are not necessarily familiar with all services and tools that are available and how they can ensure persistence within and around their research and its outputs. The researcher should be guided in formulating requirements for PID use, and when needed the data stewards should discuss these requirements with the service providers.

2.1.3.2. Assigning PIDs: by service providers

When offering services for data sharing and publication, the service provider has to make decisions about PID minting, allocation and ownership regarding master data. Service providers are in a key position in implementing FAIR by

1. assigning and managing PIDs to master data
2. integrating external PIDs and semantic artefacts in their information architecture
3. integrating external PIDs and semantic artefacts in the workflow of (meta)data creation
4. automating the processes of metadata generation and linking as much as possible in user friendly, yet transparent ways.

Services that manage research information or research outputs might reduce, enable or support the FAIR data principles (see table 1).[36]

When PIDs are created, their lifecycle should be planned and managed. The research data management services can take several different roles in the PID ecosystem as defined by the EOSC PID policy: they can be managers and owners as well as run the PID service. These responsibilities should be clear and agreed upon.

The EOSC PID policy mentions that the Kernel Information should at least contain the referent and a pointer to a type definition. These type definitions need strict curation; as there are many different ways to do data typing both in schemas as in registries and with metadata elements. PID Kernel Information recommendation done by RDA proposes a tiny amount of carefully selected metadata into a Persistent ID (PID) record (see below figure 4). This carefully chosen and placed information, targeted to internet scale services, is thought to have the potential to stimulate development of an entire ecosystem of third party services that can process billions of expected PIDs. This could be done with more information at hand about an object (no need for costly link following) than just a unique ID. The recommendation contains seven principles to enable machine actionable services.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

They state that the PID record should be a non-authoritative source for arbitrary metadata and stored directly at the resolving service.[37]

	Reduce	Enable	Support
Infrastructure level	The service does not accept persistent identifiers as values paired with natural language values when creating metadata or exposing metadata.	The service offers the option to add external persistent identifiers (DOI, ORCID) and can create new PIDs for the digital objects it hosts.	The service also offers integrated common reference metadata and presents (meta)data both for humans and machines.
Project level	The service does not enable linking versions and requires manual (free text) creation of descriptive metadata.	The service offers the possibility to create structural metadata (internal and external PIDs to versions and other relevant DOs)	The service also creates PREMIS or other types of controlled event metadata and links workflows and provenance metadata automatically

Table 1. How services can support FAIR through using PIDs.

The two-tier systems like DataCite DOI with full resolver service and landing pages have not inherently been accepted within the semantic web andand communities using LOD (Linked Open Data) and RDF: Such semantic technologies use IRIs directly as identifier and don't need the use of PIDs as an additional layer when operating with machine actionable data, see further below about RDF. Linked data solutions can be considered Deep FAIR, but they need careful management as well and also adhering to other principles like the TRUST (Transparency, Responsibility, User focus, Sustainability and Technology) principles.[38] The different use cases and contexts within the research community and sufficient nuance is necessary to meet different needs. Generally it is considered good practice to direct the human user to a landing page with metadata and licence information, when the represented object is a dataset. A persistent identifier meant for human users, for instance for data citation use, should be possible to identify as such. For example, Digital Object Identifiers (DOI) or Uniform Resource Names (URN) have distinctive namespaces that makes it easy for a researcher to recognize and use in an appropriate way. Fragments and elements might behave differently when queried by a machine or a human asking for information with a web browser. There are different ways to approach and solve these situations.

Research data is sometimes published and managed in databases, where data is published as individual nano publications and search queries might produce compiled datasets, which in turn can be given identifiers. Also queries can be stored and given persistent identifiers. This enables good prerequisites for replication and citation. In practice, dynamic and evolving datasets create challenges to implementing the FAIR principles on data. DataCite gives four alternative ways to cite dynamic datasets, which offer different levels of reproducibility:

1. Cite a specific slice or subset

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

    - the set of updates to the dataset made during a particular period of time or to a particular area of the dataset
2. Cite a specific snapshot
    - a copy of the entire dataset made at a specific time
3. Cite the continuously updated dataset, but add Access Date and Time to the citation
    - Does not necessarily ensure reproducibility
4. Cite a query, time-stamped for re-execution against a versioned database[39]

The RDA Data Citation working group[40] produced a recommendation on evolving data in 2015. The solution comprises of the following core recommendations [41]:

- Data Versioning: For retrieving earlier states of datasets, the data needs to be versioned. Markers shall indicate inserts, updates and deletes of data in the database.
- Data Timestamping: Ensure that operations on data are timestamped, i.e. any additions, deletions are marked with a timestamp.
- Data Identification: The data used shall be identified via a PID pointing to a time-stamped query, resolving to a landing page.

Another approach is nanopublication for citing parts of datasets, sometimes referred to as micro attribution.[42] This has been applied in life sciences. According to nanopub.org[43] a nanopublication is a graph with three basic elements:

1. The Assertion: An assertion is a minimal unit of thought, expressing a relationship between two concepts (called the Subject and the Object) using a third concept (called the Predicate).
2. The Provenance: This is metadata providing some context about the assertion. Provenance means, 'how this came to be' and includes the methods that were used to generate the assertion and attribution metadata such as authors, institutions, time-stamps, grants, links to DOIs, URLs about the assertion.
3. The Publication Information: This is metadata about the nanopublication as a whole, and pertains to both the assertion and provenance. Similar to the provenance graph, the Publication Information contains "citation-like" metadata but pertains to the nanopublication and not just the assertion.

Automatic metadata creation has been identified as an important way to enable FAIR data.[44] For a researcher, FAIR-born data could be the most ideal case.

2.1.4. PIDs and metadata

The EOSC Interoperability Framework mentions that PID links held in the metadata section (as pictured in Figure 4) of the FAIR digital object should resolve into the FAIR digital objects themselves in order to provide value in the ecosystem and also provide the metadata needed to support all four layers of interoperability.[45] The FAIR data principles recommend assigning PIDs both for the metadata and the data, which can create both endless recursion and confusion among users if implementation isn't clear. The RDA Data Foundation & Terminology Group[46] as well as the RDA PID Kernel Information group[37] open for the description of some of the relationships between elements within the Kernel Information of a PID, which could solve this problem. At the same time it is distinct, both in the PID KI recommendation and in the EOSC PID Policy draft[10], that the

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

authoritative source and master data is by/with the PID owner. All information added to the PID Kernel Information poses a potential risk or confusion.

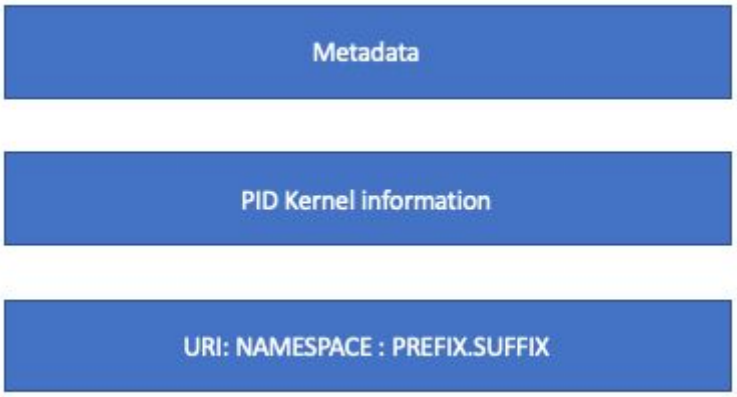

Figure 4. There are several layers of information related to the persistent identifier. There should be agreement and documentation on where information can be stored and how relations are expressed.

Concerning different types of metadata in relation to PIDs, an important distinction should be made between the PID kernel information (PID KI) in the PID record itself, which is for machines, and machines only, and other metadata about the object. The PID KI is stored in the PID record and should generally be kept as minimal as possible and contain only simple key-value pairs. Every attribute in a PID KI profile should depend only on the identified object and no other objects and should describe the object directly and not any other attribute in the same profile. When information duplicates metadata maintained elsewhere, the external source should be considered authoritative. PID systems should provide the attribute profile they support under their prefix root.[37] At its simplest the kernel metadata in the PID record only contains the PID and the referent URL, which is only one relation, but information about PID creation and owner are usually important as systematic management of the records metadata is essential for a trustworthy PID service. The PROV data model[47] was considered imperfect, but the best so far, by the FREYA project in 2019 and is now included in the PID KI recommendation.[48] Information about versioning is usually not recommended to be included in the PID record and also other relations should be stored mindfully while the master data is better stored in the metadata record. There are of course different kinds of approaches and different solutions suit different use cases depending on how the service architecture is structured. Policy, documentation and curation are the most important parts of ensuring a robust solution.

Besides provenance and information on other relations, information about the type of the digital object can be included in the PID KI. Also, when several download services are available, this information could be included, once a standard way of expression is agreed upon.

An interesting and relevant example of the structure for semantic interoperability presented by the EOSC IF (see figure 8 below) is the digitalObjectPolicy presented in the PID KI recommendation.[37] Using this property would enable registering PIDs in a flexible and trustworthy way during for instance the research process. The Object Policy property would contain its own attributes on whether the object is static or dynamic, a tombstone and also point further to a licence. It can be well argued that this kind of information should be part of the Kernel Information.[49] What could be

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

as relevant for efficient machine actionability could be standardized access information (rather than licence).

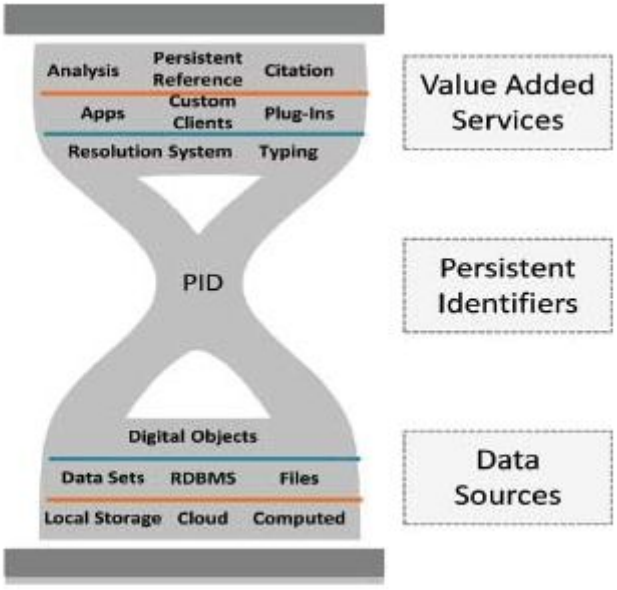

Figure 5. from Wittenburg CC-BY-4.0 2019[33] This model clearly illustrates the goal of minimizing the information carried by the PID itself as a way to gain robustness and persistence. The model assumes great alignment between all parts in the ecosystem to ensure semantic interoperability.

There is one additional aspect with classifying data objects into dynamic, static and deprecated: the classification should not be hidden and only visible to machines. It would be important for clarity, adoption and trustworthiness that the different states of the data object are visible and understandable for human users as well.

2.1.5. Research information and PIDs

There are two major dimensions of persistent identifiers in research which this question is related to. One is the management of research information and the other is the management of research data. The first is focused on metadata and serves creating knowledge about the research at large, while the latter serves scientific purposes in expressing the structures of the data and managing the lifecycle and documentation to support things like reproducibility and transparency of the research. While research information is aiming at creating as large contexts as possible on a more generic level, research data management is focussed on data management processes over time. The contexts are of course related and support each other, but still, it might be useful to separate the two use cases. As the goals are different, also the FAIR principles result in somewhat different manifestations (see table 2). The aim should still be good quality data with as little ambiguity as possible, in other words, sharing as many stable concepts (PIDs) as possible. This separation of use cases is of course not airtight, but can still help structure the discussion about PIDs and take it forward.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

	Research information	Research data
Shallow FAIR	Outputs: DOI; URN; ARK People: ORCID	Core descriptive metadata that includes necessary research information PIDs (or at least some reference metadata like controlled vocabularies), machine readable licences.
Deep FAIR	All elements expressed with persistent identifiers All relation expressed as graphs (Citations expressed as graphs)	All data elements are machine accessible and the integrity is ensured

Table 2. Inspired by Thierry Sengstag and Sofia Georgakopoulou CC-BY-4.0[50]

PIDs for other things than data are of great use also in creating Deep FAIR data. The FREYA survey on the current PID landscape 2018[51] identified several relevant object types that have mature PID infrastructure, but updated here with information from survey data and desktop research. Table 3 and 4 show some examples of proposed PID schemas and service providers behind services.

Mature contexts		
Object type	**PID schemes**	**Service providers** (not all provide proxy or landing page, but these can be used to disambiguate and link information)
Publication	DOI, Accession number, Handle, URN, Scopus EID, Web of Science UID, PMID, PMC, arXiv Identifier, BibCode, ISSN, ISBN, PURL	CrossRef, National libraries, Internet Archive
Data publication	DOI, Accession number, Handle, PURL, URN, ARK	DataCite, Internet Archive, ePIC, National libraries, B2handle
Researcher	ORCID iDs, ISNI (also DAIs, VIAFs, arxivIDs, OpenIDs, ResearcherIDs, ScopusIDs)	ORCID, National libraries/ISNI

Table 3. Contexts where mature PID systems have been identified with examples of schemas and service providers.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Emerging contexts		
Object type	**PID scheme**	**Service providers** (not all provide proxy or landing page, but these can be used to disambiguate and link information)
Organization	DOI RoR ISNI GRID Ringgold IDs Wikidata ids EU VAT numbers LEI PSI OID	DOI : Crossref (no landing page) ISNI International Agency Ltd, GRID: Digital Science solution Ltd./Holtzbrinck Publishing Group, Wikidata id: Wikimedia foundation RoR (community-led project) RIN: Ringgold Inc OID: ITU LEI: Global Legal Entity Identifier Foundation (GLEIF)
Data repositories	DOI	DataCite Elixir/ University of Oxford
Projects	local identifier, accession number, RAiD	RAiD: Australian consortium
Grants	DOI, PURL URI	DOI: CrossRef
Software	DOI, SHA-1 hash, SWH, commit hash	Zenodo/DataCite, Software Heritage
Instrument, Device, Sensor, Platform, Research Facility	DOI, RRID, UUID	DataCite RRID: University of California DOI: Journal of large-scale research facilities JLSRF
Field Station	deims id	DEIMS
Physical Sample or object	Accession number, RRID, DOI, IGSN, URN	RRID: University of California

Table 4. Contexts where emerging PID systems have been identified with examples of schemas and service providers.

For creating deep FAIR solutions the input of service providers and developers is important. The service providers should aim at creating solutions that utilize trustworthy and suitable PIDs for each use case and be mindful about the management of PIDs.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2.2. Enabling FAIR with semantic interoperability

2.2.1. Enabling FAIR with semantic interoperability: by researchers

Interoperability, the third letter of FAIR, is commonly considered the most difficult one to achieve. Proper interoperability can save a lot of time in research because it minimises misunderstandings and also makes it much easier to combine datasets coming from different sources: either similar datasets joined together to make a larger study, or dissimilar datasets that are joined together to solve complex problems.

Interoperability needs to address both the format in which data is stored, and the description of the data (all the data elements) in a way that makes sure that it can be interpreted by others, even in 10 years. This requires the use of open standards for data formats, flexible enough to accommodate the variety of data expected, and on the other hand documented enough that every aspect of the data is clarified without leaving any ambiguity. This makes it necessary e.g. that all values are associated with data types and numeric values have explicit units, and the meaning of a missing or an out of range data value is defined. It also requires that relations between data items are explicit (e.g. a "temperature" is the temperature of the given "object", and not of the environment).

2.2.2. Enabling FAIR with semantic interoperability: by data stewards

2.2.2.1. The issue being addressed

The FAIR in Practice Taskforce of the EOSC FAIR Working group identified[7] that interoperability is the hardest of the four letters to attain. In this section we will give some consideration to interoperability in the FAIR context and techniques that have been successfully used to implement it.

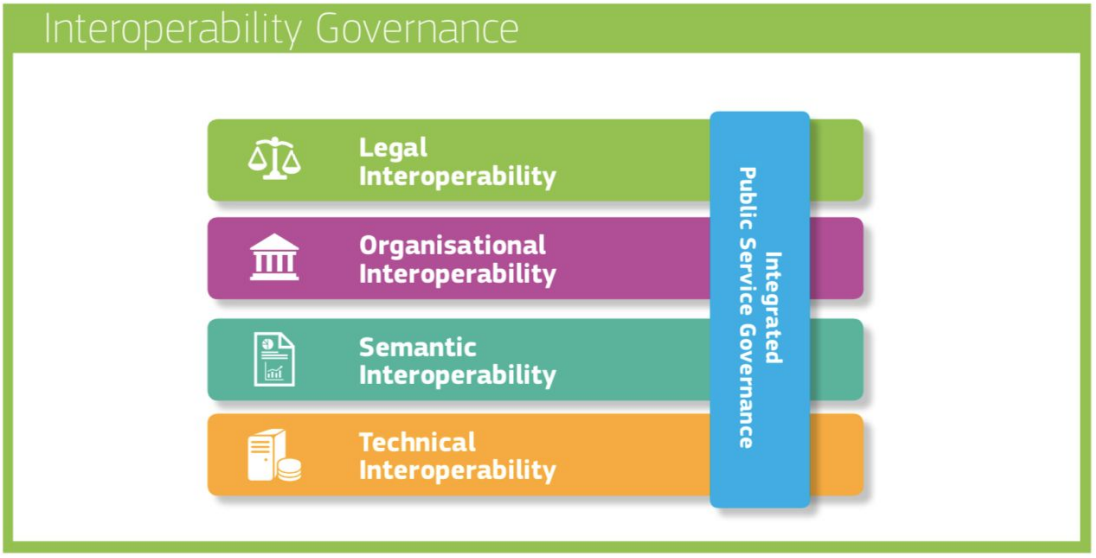

Figure 6. The European Interoperability Framework promotes seamless services and data flows for European public administrations[45]

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

In Figure 6, four levels of interoperability are shown as they are defined in the European Interoperability Framework.[52] The EOSC interoperability framework[45] describes these as:

Technical interoperability: the ability of different information technology systems and software applications to communicate and exchange data.

Semantic interoperability: the ability of computer systems to exchange data with unambiguous, shared meaning.[53]

Organisational interoperability: the way in which organisations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals.

Legal interoperability: the broader environment of laws, policies, procedures and cooperation agreements needed to allow the seamless exchange of information between different organisations, regions and countries.

The FAIR principles focus on requirements for semantic and technical interoperability: technical interoperability requires e.g. standardization of file formats, whereas semantic interoperability addresses the expression of the (meta)data inside. Prof. Barend Mons summarizes semantic interoperability as "The machine knows what I mean".[54]

2.2.2.2. What is semantic interoperability trying to do?

An important application of interoperability for research data is to be able to bring data from different sources together without having to go through an extensive round of re-interpretation. One can recognize that there are various grades of difficulty for this problem:

1. The easiest is integration of data that comes from two successive similar projects by the same researcher. Interoperability at this level requires that the same definitions are used for the same fields, and the same data file formats. This basically requires that the researcher does not change his ways.

2. The next level of complexity is when integrating data from two researchers doing similar projects in the same lab. Now, they need to make sure that they use the same data standards.

3. One level up is integration of similar data from different institutions. Again, the same standards should be used, but there is more chance of misunderstanding.

4. The final level of data integration is interdisciplinary data integration, bringing together data from different (sub)fields of research and/or society.

With each successive level, avoiding misunderstandings requires greater precision in documentation. For example, among meteorologists the data type "temperature" is rarely misunderstood, but when their data needs to be integrated with medical data it becomes essential to distinguish "body temperature" from "ambient temperature". Especially at level 3 and 4, in some cases differences in human language used to express the data can also complicate interoperability.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

Without special care for interoperability, it is very common that extensive work on re-interpretation is needed, when different datasets are integrated: research has shown that data pre-processing, also called data wrangling or data munging[55], often requires 75% of the time used for data processing.[56] This also means there is a huge potential here for saving time on data processing, if assuring semantic interoperability by the complete research community is becoming common practice.

If we want the "machine to know what we mean" we need to clear a few hurdles:

1. We should make sure that when two data files are "talking about the same thing", they ideally use the same way to do that. If that is not possible, we need to be explicit that two different terms or identifiers are the same. There are deeper difficulties with this that are out of the scope of this document to address: researchers may differ in opinion about whether two things are "the same" or not, depending on the context of what they are researching. See e.g.[57]
2. We should make sure that it is obvious when two data files are "talking about different things", i.e. we should avoid ambiguity.
3. We need to make sure to explicitly describe every anomaly that could be present in the data. For example, it should be documented whether a missing data value means that the value has been determined to be empty, that it has not been measured or that its value is unknown or irrelevant.
4. We need to make sure that no guesswork is needed to interpret relationships between data items. For example, in a data table listing: patients, their illness, and medication they are taking. A human user can *assume* that the medication is taken in an attempt to cure or mitigate the illness; semantic interoperability demands that this relation is explicitly described, so a machine can also use it.

To do this, systems for semantic interoperability do the following:

1. Rather than using a *term* to describe something that is specific to a human language and may be context-dependent in practice, they prefer to refer to *concepts* that represented with by a unique *identifier*.[58]
2. Each data value is associated with a precise *data type*, documented to such precision that misunderstandings are avoided.

2.2.2.3. Technology supporting semantic interoperability

The original paper describing the FAIR principles carefully avoided choosing a single technology to implement them. It can be seen, however, that research communities that are currently the furthest along the road to implement interoperability for FAIR data have often chosen Linked Data[59] technologies (recognized by abbreviations like RDF, OWL, SKOS, SPARQL). This is not because this solution is the only approach that could be taken, but because it has been under development for many years and is currently one of the few frameworks that can make data unambiguously understandable. It is possible that the role played by Linked Data today could be played by another more powerful approach in the future[56], but Linked Data has the best potential for implementing FAIR interoperability at this moment.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

The core of Linked Data (see Figure 7) is formed by the Resource Description Framework (RDF).[60] All data represented in RDF is turned into so-called "semantic triples" that each contain a statement in the form of subject-predicate-object, each describing either a property of something (subject: "bird" predicate: "body temperature in degrees C" object: "41") or a relation between two things (subject: "Alice" predicate: "Talks to" object: "Bob"). In order to make these statements machine readable in RDF, subjects and predicates are never just a character string, but always a Internationalized Resource Identifier (IRI).[61] The objects in an RDF triple can either be a IRI, literal or blank node, and if it is a literal then there are techniques to make sure it is well described; e.g. the language of a string is explicitly added. To represent knowledge, multiple RDF triples are put together in a network or "graph".

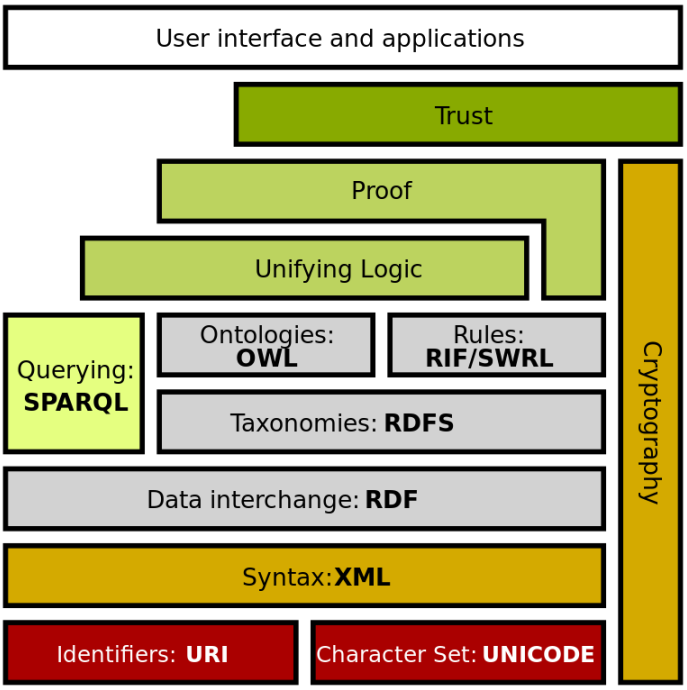

Figure 7: Semantic Web Stack[62]

IRIs have a central place as unique identifiers for subjects, predicates and objects in RDF. In the design of the Linked Data paradigm, a central assumption is that these IRIs are the unique identifiers of the concepts in the graph, and it is good practice to use IRIs that correspond to human-readable and/or machine readable explanations of the concepts. IRIs, however, do not satisfy all requirements that are currently expected from PIDs, as described in the associated sections 2.1.3 Creating sustainable interoperability with PIDs and 2.1.3.2. Assigning PIDs : by service providers of this report. It is, however, often quite easy to use PIDs when using Linked Data by replacing regular IRIs by IRIs that are formed by combining PIDs with their resolver on the World Wide Web.

When using Linked Data, the IRIs for related things are stored in collections, which are named differently depending on the structure and how the relationships are described. They can be named as glossary, controlled vocabulary, thesaurus, ontology or data model. All of these together are

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

called “semantic artefacts”. In FAIRsFAIR Task 2.2 our preference for this term (over “ontology”) is explained.[63]

Data represented in RDF differs from data that is traditionally represented in tables in two important ways:

1. RDF is relatively flexible: in contrast to objects described in a table, it is not required that each object described in RDF has values for the same set of properties. This makes it equally easy in RDF to describe one-to-one relationships as one-to-many relationships, whereas in tabular data one-to-one relationships can be expressed using columns in a table, but one-to-many relationships require the creation of multiple tables with explicit connections.

2. RDF assumes an “open world”[64]: In contrast to objects described in a table where an “empty box” could mean that the object does not have the corresponding property, in RDF the absence of a triple describing a property (as a simple example: an owner) can only mean that it is *unknown*. Someone else may describe that property elsewhere. If it is *known* that a value does *not* exist (e.g. we know that the object does not have an owner), that is a property that should be made explicit. The open world makes it very easy for anyone to build upon existing knowledge in RDF, whereas it is virtually impossible to “add a column” to an existing data table or “fill in empty values” for anyone except the maintainer of an existing tabular dataset.

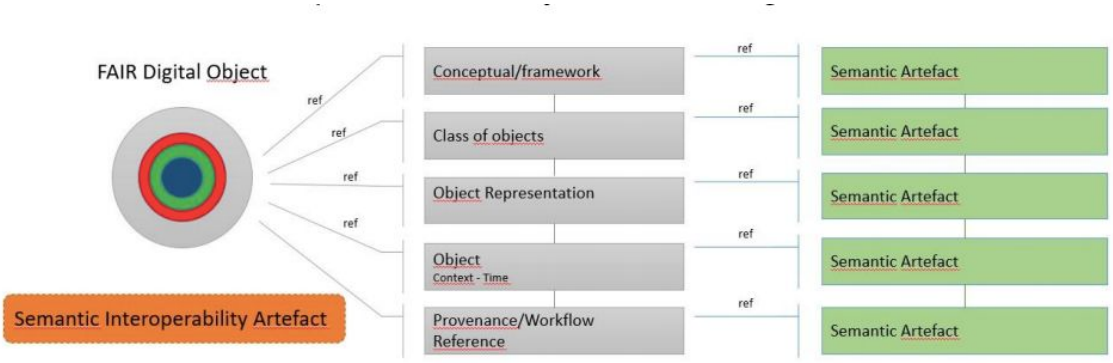

Figure 8. A model presented in the EOSC IF draft of a structure that supports semantic interoperability by including rich semantic references.[42]

Many of the complexities in the implementation of RDF arise from its flexibility. Software that uses the data often makes assumptions about a data structure. Tabular data makes this easy because it enforces this structure. For RDF, a structure is more difficult to guarantee, and data structure validation is provided separately. One modern example of a technique used for this is the Shapes Constraint Language[65] (SHACL). SHACL writes simple statements (in RDF!) about data that describes the structure of that data. It allows matching a *data graph* with a *shapes graph*. This way, SHACL makes it possible to describe what properties and relationships the nodes in the graph must have and must not have, use this to filter the graph, and raise an error when conditions are not met. If there is no error, software querying the graph can be sure that it satisfies the structure it needs.

There are different ways in which RDF data can be represented. The original RDF was an XML markup language that is very voluminous and hard to understand for humans. An RDF graph can also be represented in other ways, of which TTL (pronounced “Turtle” and easier to read by

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

humans), and JSON-LD are popular. Conversion between the different representations is simple and does not lose any information, so the variety of formats does not introduce additional semantic interoperability problems.

A common misconception about Linked Data is that it requires all data to be represented as RDF. In fact, RDF is a very inefficient way of storing and processing many forms of dense data. The requirement that is placed upon us for interoperability is not that the data is all converted to RDF, but that someone *skilled in the field* would be able to set up an automated process to perform that conversion fully automatically. However, it should be noticed that if data must still be interoperable in 10 years, that either requires that there will still be people available at that time who are *skilled in the field*, or, alternatively, that the automated process for the conversion is actually built. In practice, dense tabular research data is often left as such, and combined with abstracted conclusions and metadata represented as RDF, see for example the Allotrope Framework.[66]

2.2.2.4. Semantic artefacts

In the previous report the availability of good semantic artefacts was considered as one of the useful ways to achieve good quality metadata and promote FAIRness. Deliverables from FAIRsFAIR Task 2.2 describe recommendations and good practices for the creation and maintenance of semantic artefacts.[67] The authors also facilitated a new RDA task group to propose a common minimal metadata schema for these. The RDA Vocabulary Services IG also does highly relevant work on FAIR semantic artefacts.[68,69] According to the EOSC Interoperability Framework there should be specific focus on the management and governance of semantic artefacts and on ensuring their FAIR properties.

The use of semantic artefacts should be integrated in the workflow in ways that make it possible to create interoperable (meta)data. Semantic artefacts are tied to the research domain that uses them, and this creates a large context diversity. This is, however, not the only diversity that needs to be considered: a challenging area in semantic interoperability is cross-language interoperability (cultural and linguistic), which requires multilingual semantic artefacts (eg. vocabularies, ontologies and concept schemes having terms in different EU languages). This is a dimension that is especially important for humanities and social sciences, but should be considered as a generic point because it is important for findability (one cannot search for something in a language that one does not write), interoperability (it enables joining local data sets together internationally), and generally reusability, open science, societal impact and outreach. Even though English is used as a standard language in research, not all researchers are fluent. Additionally, inviting broader contributions (citizen science) and communicating results both benefit from research tools being available in local languages. Inter-language interoperability is an issue where Europe can use its cultural diversity to turn a challenge into a possibility and develop scientific resources that are available and reusable for a larger audience. The EU terminology could also be developed by linking or extending it to the terminology of the research domain and EOSC.[70]

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

2.3. Enabling FAIR with metadata

Actress Jean Harlow famously said "Don't give me books for Christmas; I already have a book".[71] This section tries to resolve a similar misunderstanding of the variability of "metadata": more than one kind of metadata is required in order to make data FAIR.

Modeling and metadata need significant attention. Sometimes metadata is missing, and when available, it may not give sufficient information about a dataset for a re-user. 'Search-metadata' needs sufficient detail to be able to distinguish between sets that are useful and sets that are not useful for the searcher. There is a need for examples of what FAIR metadata looks like and practical guidelines on improving FAIRness. It's important to minimize the burden on people to provide the metadata, and maximize the benefits.

Note that there are upcoming deliverables in FAIRsFAIR that address metadata interoperability, especially D3.6 *Proposal on integration of metadata catalogues to support cross-disciplinary FAIR uptake* and D3.7 *Report on integration of metadata catalogues*, as well as the reference implementation done in WP2 task 2.3. Here we focus mainly on the role metadata has in creating semantic interoperability.

2.3.1. Enabling FAIR with metadata: by researchers

Several of the FAIR principles call for the presence of *metadata*. This causes confusion because the term is very abstract and having one kind of metadata does not necessarily satisfy the needs of any of the other FAIR principles. Metadata that describes things like authorship, time, date is the only type of metadata that is universally applicable, and one of its oldest standardizations, Dublin Core[72], is thus mentioned in many places as a prototypical example of metadata. This sometimes leads to the misunderstanding that providing these metadata items would be sufficient. More than generally describing "where the data comes from", metadata should provide the *documentation* of the data needed not only for visibility and citation, but also to ensure reusability and offer sufficient evidence for research.

Note also that each metadata standard has two separate components. First, there is the schema of metadata items, often separated into subsets of "obligatory", "recommended" and "optional" items. Second there is the way in which each item must be specified. This strongly determines how well each metadata standard satisfies the needs for semantic interoperability. E.g. since Dublin Core metadata contains free text entries, it does not automatically imply semantic interoperability. Modern metadata standards require that values are picked from specific semantic artefacts (ontologies and vocabularies, see the Semantic artefacts section), this restriction can make it more cumbersome to assemble the metadata, but the resulting metadata will be a much better facilitator of interoperability.

Six of the 15 FAIR principles mention metadata in a way that is relevant here. We will cite and interpret each of these in order.

FAIR **principle F4**, *(Meta)data are registered or indexed in a searchable resource*, refers to metadata that describe the identity of the data. This can be metadata about where it comes from (creator, date, etc, coming from the Dublin Core Metadata Items (DCMI) as described above or from DataCite[39], but should also include other things that we can expect others who want to reuse the

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

data to search for, such as subject-area, topics and keywords. Note that many of the obligatory DCMI fields are "free text", which does not make it interoperable: for example topics selected from an ontology make it much easier to find a dataset (but take work to add to the metadata) than free text keywords (which are arguably much easier to add). This kind of metadata is not only used for searching, but also for attribution and recognition of contributions. One can not only express authorship, but a whole range of different types of contributions like curation and collection of data using metadata following DataCite and the CREDIT taxonomy[73]. Adding these is a valuable opportunity to make the work of various people involved visible.

FAIR **principle I1**, *(Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation*. This is something that has to be done in cooperation with the data steward and service provider. The researcher should be sure to consult data stewards to scan the landscape of possible solutions and discuss how these can best be reused and implemented.

FAIR **principle I2**, *(Meta)data use vocabularies that follow FAIR principles.* The data elements should come from FAIR semantic artefacts whenever possible. The researcher is dependent on available services and semantic artefacts to encode the metadata. Researchers should also engage in terminology and ontology work by providing information and definitions when structured semantic artefacts are created. A data steward can support both processes. Schemas and application profiles should be published in machine readable formats. Terminology work (as in defining scientific concepts and terms) should be considered scientific work (and thus, also rewarded and appreciated as such).

FAIR **principle R1.1**, *(Meta)data are released with a clear and accessible data usage license.* Not only humans, but also machines should ideally have the possibility of judging whether data are available for a particular type of re-use, for example whether it will be compatible with consent of a data subject. This requires that the rights are encoded in an interoperable way. A researcher, together with his data-steward, can assure that rights management is done from the beginning of the research process to ensure later implementations of FAIR: Machine readable access and licence information should be used where possible. Controlled values and schemas should be used as widely as possible. Technical implementation of rights management requires a well managed Authentication and Authorisation Infrastructure (AAI) system and clear policy alignment (i.e. technical, organisational and legal interoperability).

FAIR **principle R1.2**, *(Meta)data are associated with detailed provenance.* This means not only information about creators, source data, research questions or projects and other contextual information that is relevant for assessing possible reuse, but also information about data lifecycle events and technical information about used software, protocols and methods.[74]

FAIR **principle R1.3**, *(Meta)data meet domain-relevant community standard*s. The domain-relevant metadata referred to here are often so-called "minimal information" standards, for which the abbreviated name starts with "MI(A)" for "minimal information (about) ... ". Where domain relevant community standards are missing, steps should be taken to create these both regarding schemas and vocabularies as the development of semantic artifacts if needed. The community should do this together with the data steward to ensure that you can reuse as many existing resources as possible.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

It is likely that for each of these metadata types, different metadata standards are needed. Some of these are specific to the research discipline. It is important to realize that it is good to find out which metadata is needed as early as possible in a project: this will make it possible to collect the right metadata from the start, which will be much less time consuming than collecting it at the end of the project, the last moment the data can be annotated and preserved.

2.3.2. Enabling FAIR with metadata: by data stewards

There are different kinds of tools for choosing and compiling metadata schemas, like FAIRsharing[75], the DCC metadata catalogue[76], the RDA metadata directory[77] or component libraries like CEDAR[78] or CMDI.[79] If a suitable metadata schema does not exist, the first choice should be to combine elements from existing schemas. This is a better option than coming up with something completely self-invented. It is, for instance, a good idea to take a common format or vocabulary as a base, like DCAT[80], DDI[81] or DataCite[39] and extend it with elements from other schemas or vocabularies where necessary. Community-specific metadata schemes are best created with the involvement from a significant group of researchers from that community: this will lead to better and more complete standards, and will also help adoption since those involved will have an incentive to start using the standard themselves.

Metadata schemas should be linked to semantic artefacts. This is sometimes expressed as the aim for the (meta)data to be “formal syntax and declared semantics”.[82] It’s usually not enough to name a metadata schema: it needs to be accompanied by more specific documentation about the metadata elements and how these are used. An elaborate way of doing this is described in the EOSC Interoperability Framework (chapter 4).[45]

Schemas and application profiles[83] should be published in machine readable formats. It is important to note that a metadata schema is often not enough, especially when we want to create domain agnostic metadata. Therefore application profiles are necessary specification, as much specification needs to be done explicitly. It is also a good idea to look at the FAIRsFAIR Data Object Assessment Metrics[9] when planning and creating metadata. This can help in making choices that support FAIR data.

The data steward needs to bring together the requirements and ontology (the concepts, the terms and their relations) of the scientific field and facilitate the researcher in finding the correct forms to represent their knowledge, and that documentation is sufficient. The data steward also needs to make the service provider aware of the users needs.

It is also important that the data steward engages researchers in using and developing semantic artefacts.

3. Discussion and conclusions

A generic solution for achieving FAIRness does not exist. Every case requires a careful consideration of the investments needed to make data more FAIR and beneficial for the scientific community. This consideration requires up-to-date knowledge of the available technologies. In this report, we have chosen to document three components of FAIR where we see possibilities for broader adoption and convergence: persistent identifiers, semantic interoperability using linked data, and metadata. By providing this information in separate sections targeted towards researchers, data

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

stewards and service providers, we have attempted to encourage good practice with actionable guidance.

Connections between the sections: The content in each section of the report has been organised in alignment with the needs of each target audience - with information for data stewards and service providers containing increasing levels of detail. In order to know what their customers will be expecting, there is value for data stewards in the sections targeted for researchers, and there is value for service providers in reading what is written for data stewards.

Conclusions for Researchers: Aim at consistent use of PIDs, supported by good metadata. This will enable findability of your research outputs and make research data management easier in the long run. In order to increase interoperability, try to imagine making sense of your data in 10 years: is there any "implicit" knowledge, e.g. about data types, that can be made explicit? Before creating datasets, plan the use of PIDs, data formats, and metadata with help and guidance from your data stewards. In other words, don't try to do this alone - be sure to contact your data stewards for support in developing sustainable PIDs and metadata, thus increasing FAIRness.

Conclusions for Data stewards: It is important to think about interoperability and longevity. Support researchers in determining the appropriate depth of FAIRness of the data and provide them with examples of what good FAIR PIDs and metadata should look like. In addition, educate researchers about reproducibility and semantic artefacts. Think with the researchers about the understandability of the data in 10 years; making assumptions on explicit data documentation will really help interoperability. Engaging researchers in related discussions, and decision-making processes would help to improve FAIRness from the grass roots.

There are many (good) solutions for developing FAIRness. We encourage data stewards to use existing services, instead of trying to implement their own services from scratch.

Conclusions for Service providers: Researchers and data stewards alike need the service providers' support in making appropriate use of the solutions for implementing FAIR. Consider each need and use case by evaluating the FAIR principles and assessing the value of implementing them. We recommend conducting a cost-benefit analysis on each principle with a sustainability perspective: what can be managed and curated over time? Strive to support scientific reproducibility and data lifecycle management with well documented technologies, well managed services and workflows, and curated data.

Pursuing interoperability requires a relatively large investment compared to the other FAIR aspects, but it has promising benefits and should not be overlooked. Service providers can support data stewards (and consequently, researchers) in selecting the technology suitable for implementing the required level of semantic interoperability within a specific scientific context.

We acknowledge that while technology is a tool for solving many 'technical' problems, the bigger challenge lies in the human element: humans need to understand how they can help each other and machines to create findable, accessible, interoperable and reusable data and research outputs. Therefore, the most important action is **co-creation and co-development** - with researchers, data stewards and service providers working together to improve FAIRness in practice. The needs of the

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

designated scientific society should always be at the core of any solution. The following quote describes a principle that applies to multistakeholder, collaborative development of FAIR:

> *Always design a thing by considering it in its next larger context – a chair in a room, a room in a house, a house in an environment, an environment in a city plan.*
>
> — Eliel Saarinen, Finnish architect (1873--1950)

3.1. The art of misunderstanding

Semantic interoperability and metadata are geared towards understanding data, both for humans and machines. As described, more and more well defined metadata is required to enable people who are more distant to the data to understand it: both in subject matter (for interdisciplinary interoperability) and over time (to enable long time preservation).

One especially interesting technique that can be used to verify that metadata is sufficient is to set up somebody with the task to try and deliberately misinterpret the metadata: to try and find alternative interpretations. Any successful deliberate misinterpretations are a sign of ambiguities. Removing such ambiguities can be used to sharpen the definition of the metadata that is accompanying a dataset and this will save a lot of time when the data is actually reused.

4. Acknowledgements

We would like to thank the following people for providing valuable input through comments and discussions at different stages during the writing of this report (with the authors being responsible for all possible shortcomings, errors and flaws in the text):

Rene van Horik, Miika Tuisku, Sophie Aubin, Robert Giessmann, Emilie Blotiere, Joe Tevis, Judith Pijnacker, Ronald Cornet, Martin Matthiesen, Gerard Coen, Mark Portier, Xiaoyu Fang, Frances Madden, Ulrich Schwardmann, Brian Matthews, Robert Ulrich & Jerry de Vries.

5. Bibliography

1. Lehväslaiho H, Parland-von Essen J, Behnke C, et al. D2.1 Report on FAIR requirements for persistence and interoperability 2019. Published online November 29, 2019. doi:10.5281/zenodo.3557381
2. Persistence and interoperability in FAIR research data management | FAIRsFAIR. Accessed July 29, 2020. https://fairsfair.eu/events/persistence-and-interoperability-fair-research-data-management
3. FAIRsFAIR documents for community review | FAIRsFAIR. Accessed July 29, 2020. https://fairsfair.eu/fairsfair-deliverables-community-review
4. Persistent Identifiers and Interoperability: Outcomes from the FAIRsFAIR Survey of the European Scientific Data Landscape | FAIRsFAIR. Accessed July 29, 2020. https://fairsfair.eu/news/persistent-identifiers-and-interoperability-outcomes-fairsfair-survey-european-scientific-data
5. Feedback and input on FAIR requirements for persistence and interoperability | FAIRsFAIR. Accessed July 29, 2020. https://fairsfair.eu/events/feedback-and-input-fair-requirements-persistence-and-interoperability
6. European Strategy Forum on Research Infrastructures (ESFRI) | European Commission. Accessed July 29, 2020.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

https://ec.europa.eu/info/research-and-innovation/strategy/european-research-infrastructures/esfri_en
7. Deniz Beyan O, Chue Hong N, Cozzini S, et al. *Seven Recommendations for Implementation of FAIR Practice*. Zenodo; 2020. doi:10.5281/zenodo.3931993
8. TeD-T, the Term Definition Tool of the Data Foundation and Terminology Interest Group (DFT IG) of the Research Data Alliance (RDA). Vide FAIR Digital Objects. Accessed October 3, 2019. Was available from: https://smw-rda.esc.rzg.mpg.de/index.php?title=FAIR_Digital_Objects
9. Devaraju A, Huber R, Mokrane M, et al. *FAIRsFAIR Data Object Assessment Metrics*. Zenodo; 2020. doi:10.5281/zenodo.3934401
10. Hellström M, Heughebaert A, Kotarski R, et al. Second draft Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC). Published online May 1, 2020. doi:10.5281/zenodo.3780423
11. Wittenburg P, Hellström M, Zwölf C-M, et al. *Persistent Identifiers: Consolidated Assertions. Status of November, 2017.* Zenodo; 2017. doi:10.5281/zenodo.1116189
12. Wimalaratne S, Fenner M. D2.1 PID Resolution Services Best Practices. Published online June 25, 2018. doi:10.5281/zenodo.1324300
13. Sicilia M-A, García-Barriocanal E, Sánchez-Alonso S, Cuadrado J-J. Decentralized Persistent Identifiers: a basic model for immutable handlers. *Procedia Comput Sci*. 2019;146:123-130. doi:10.1016/j.procs.2019.01.087
14. Robert E. Kahn, Christophe Blanchi, Laurence Lannom, et al. Digital object interface protocol specification. Accessed July 29, 2020. https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf
15. ITU. *X.1255 : Framework for Discovery of Identity Management Information. Recommendation X.1255 (09/13)*. ITU; 2013. https://www.itu.int/rec/T-REC-X.1255-201309-I.
16. Knowledge Hub. The PID Forum. Accessed August 6, 2020. https://www.pidforum.org/c/knowledge-hub/11
17. Link rot. In: *Wikipedia*. ; 2020. Accessed July 29, 2020. https://en.wikipedia.org/w/index.php?title=Link_rot&oldid=965642598
18. Hervé L'Hours, Ilona von Stein. *FAIR Ecosystem Components: Vision*. Zenodo; 2020. doi:10.5281/zenodo.3734273
19. Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LOB, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Inf Serv Use*. 2017;37(1):49-56. doi:10.3233/ISU-170824
20. Lasser J. Creating an executable paper is a journey through Open Science. *Commun Phys*. 2020;3(1):1-5. doi:10.1038/s42005-020-00403-4
21. EOSC-Synergy Software Quality Assurance – EOSC synergy. Accessed July 29, 2020. https://www.eosc-synergy.eu/eosc-synergy-software-quality-assurance/
22. librarycarpentry. Research Software. Top 10 FAIR Data & Software Things. doi:10.5281/zenodo.2555498
23. Discover Everything - protocols.io. Accessed July 29, 2020. https://www.protocols.io/
24. Amstutz P, Crusoe MR, Tijanić N, et al. Common Workflow Language, v1.0. Published online July 8, 2016. doi:10.6084/m9.figshare.3115156.v2
25. ro-crate. ro-crate. Accessed July 29, 2020. http://www.researchobject.org/ro-crate/
26. Research Activity Identifier (RAiD). raid. Accessed July 29, 2020. https://www.raid.org.au
27. PID Services Registry. Accessed August 6, 2020. https://www.pidservices.org/
28. Data Type Registry. Accessed July 29, 2020. http://typeregistry.org/registrar/#
29. Persistent Identification of Instruments WG. RDA. Published July 28, 2017. Accessed July 29, 2020. https://www.rd-alliance.org/groups/persistent-identification-instruments-wg
30. *Rdawg-Pidinst/Schema*. RDA WG Persistent Identification of Instruments; 2020. Accessed July 29, 2020. https://github.com/rdawg-pidinst/schema
31. Stocker M, Darroch L, Krahl R, et al. Persistent Identification of Instruments. *Data Sci J*. 2020;19(1):18.

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

doi:10.5334/dsj-2020-018
32. Juty N, Wimalaratne SM, Soiland-Reyes S, Kunze J, Goble CA, Clark T. Unique, Persistent, Resolvable: Identifiers as the Foundation of FAIR. *Data Intell*. 2019;2(1-2):30-39. doi:10.1162/dint_a_00025
33. Wittenburg P. From Persistent Identifiers to Digital Objects to Make Data Science More Efficient. *Data Intell*. 2019;1(1):6-21. doi:10.1162/dint_a_00004
34. Compact Identifiers. Accessed July 29, 2020. https://n2t.net/e/compact_ids.html
35. Identifiers.org. Accessed July 14, 2020. https://docs.identifiers.org/
36. Koers H, Gruenpeter M, Herterich P, et al. Assessment report on "FAIRness of services." Published online February 28, 2020. doi:10.5281/zenodo.3688762
37. Weigel T, Plale B, Parsons M, et al. RDA Recommendation on PID Kernel Information. Published online 2018. doi:10.15497/RDA00031
38. The TRUST Principles - An RDA Community Effort. RDA. Published May 18, 2020. Accessed July 29, 2020. https://www.rd-alliance.org/trust-principles-rda-community-effort
39. DataCite Schema. DataCite Schema. Accessed July 29, 2020. https://schema.datacite.org/meta/kernel-4.1/index.html
40. Data Citation WG | RDA. Accessed July 29, 2020. https://www.rd-alliance.org/groups/data-citation-wg.html
41. Rauber A, Asmi A, van Uytvanck D, Proell S. Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). Published online October 20, 2015. doi:10.15497/RDA00016
42. Fabris E, Kuhn T, Silvello G. A Framework for Citing Nanopublications. In: *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Proceedings*. Springer Verlag; 2019:70-83. doi:10.1007/978-3-030-30760-8_6
43. Nanopublications. Accessed July 29, 2020. http://nanopub.org/wordpress/
44. Koers H, Bangert D, Hermans E, Horik R van, Jong M de, Mokrane M. Recommendations for Services in a FAIR Data Ecosystem. *Patterns*. 2020;0(0). doi:10.1016/j.patter.2020.100058
45. Corcho O, Eriksson M, Kurowski K, et al. *EOSC Interoperability Framework : DRAFT*.; 2020. https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf
46. Berg-Cross G, Ritz R, Wittenburg P. RDA DFT Core Terms and Model. Published 2016. Accessed July 14, 2020. http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318
47. PROV-DM: The PROV Data Model. Accessed July 29, 2020. https://www.w3.org/TR/prov-dm/
48. Martin Fenner, Joe Wass, Tom Demeranville, Sarala Wimalaratne, Richard Hallett. D2.2 PID Metadata Provenance. Published online June 18, 2019. doi:10.5281/zenodo.3248653
49. Parland-von Essen J, Fält K, Maalick Z, Alonen M, Gonzalez E. Supporting FAIR data: categorization of research data as a tool in data management. *Informaatiotutkimus*. 2018;37(4). doi:10.23978/inf.77419
50. funding FAIR communities a proposal by the Research Data Management network of University of Basel. https://researchdata.unibas.ch/en/home/. Accessed August 26, 2020. https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK_P-2/2019_OpenScience_Poster12_RDM_UniBas_20190909_fundingFAIRcommunities.pdf
51. Christine Ferguson, Jo McEntrye, Vasily Bunakov, et al. D3.1Survey of Current PID Services Landscape - Revised. Published online October 18, 2019. doi:10.5281/zenodo.3554255
52. European Union Directorate-General for Informatics. *New European Interoperability Framework - Promoting Seamless Services and Data Flows for European Public Administrations*. European Commission; 2017. doi:10.2799/78681
53. Semantic interoperability. In: *Wikipedia*. ; 2020. Accessed July 29, 2020. https://en.wikipedia.org/w/index.php?title=Semantic_interoperability&oldid=970122164
54. Lex Nederbragt op Twitter: "Listening to Barend Mons @barendmons on GoFair @Realfagsbibl: 'If I am asked to summarise the FAIR principles, I say: "The machine knows what I mean"' @GOFAIRofficial https://t.co/OhJFgtfw4M #gofair" / Twitter. Twitter. Accessed July 29, 2020. https://twitter.com/lexnederbragt/status/1070676797023576064

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

55. Data wrangling. In: *Wikipedia*. ; 2020. Accessed July 29, 2020. https://en.wikipedia.org/w/index.php?title=Data_wrangling&oldid=966597395
56. Wittenburg P, Strawn G, Mons B, Boninho L, Schultes E. Digital Objects as Drivers towards Convergence in Data Infrastructures. Published online December 2018. doi:http://doi.org/10.23728/b2share.b605d85809ca45679b110719b6c6cb11
57. Batchelor C, Brenninkmeijer CYA, Chichester C, et al. Scientific Lenses to Support Multiple Views over Linked Chemistry Data. In: *The Semantic Web – ISWC 2014*. Springer, Cham; 2014:98-113. doi:10.1007/978-3-319-11964-9_7
58. Triangle of reference. In: *Wikipedia*. ; 2019. Accessed August 20, 2020. https://en.wikipedia.org/w/index.php?title=Triangle_of_reference&oldid=895514020
59. Linked data. In: *Wikipedia*. ; 2020. Accessed July 14, 2020. https://en.wikipedia.org/w/index.php?title=Linked_data&oldid=965913447
60. Resource Description Framework. In: *Wikipedia*. ; 2020. Accessed July 14, 2020. https://en.wikipedia.org/w/index.php?title=Resource_Description_Framework&oldid=956820853
61. Internationalized Resource Identifier. In: *Wikipedia*. ; 2020. Accessed August 26, 2020. https://en.wikipedia.org/w/index.php?title=Internationalized_Resource_Identifier&oldid=936752293
62. Semantic Web Stack. In: *Wikipedia*. ; 2019. Accessed August 20, 2020. https://en.wikipedia.org/w/index.php?title=Semantic_Web_Stack&oldid=922083475
63. Coen G. Introduction to Semantic Artefacts. Presented at the: October 22, 2019. doi:10.5281/zenodo.3549375
64. Open-world assumption. In: *Wikipedia*. ; 2019. Accessed July 14, 2020. https://en.wikipedia.org/w/index.php?title=Open-world_assumption&oldid=922853720
65. Shapes Constraint Language (SHACL) W3C Recommendation. Published July 20, 2017. Accessed November 22, 2019. https://www.w3.org/TR/shacl/
66. The Allotrope Framework. allotropefoundation. Accessed July 30, 2020. https://www.allotrope.org/allotrope-framework
67. Le Franc Y, Parland-von Essen J, Bonino L, Lehväslaiho H, Coen G, Staiger C. D2.2 FAIR Semantics: First recommendations. Published online March 12, 2020. doi:10.5281/zenodo.3707985
68. Vocabulary Services IG. RDA. Published March 10, 2015. Accessed July 30, 2020. https://www.rd-alliance.org/groups/vocabulary-services-interest-group.html
69. VSIG/VSSIG re-configuration. RDA. Published June 27, 2017. Accessed July 30, 2020. https://www.rd-alliance.org/group/vocabulary-services-interest-group/post/vsigvssig-re-configuration
70. IATE - Search. Accessed July 30, 2020. https://iate.europa.eu/home
71. Jean Harlow Quotes. BrainyQuote. Accessed July 30, 2020. https://www.brainyquote.com/quotes/jean_harlow_539678
72. DCMI: Home. Accessed July 30, 2020. https://www.dublincore.org/
73. Search Results for “contributor roles” – CRediT. Accessed July 30, 2020. http://credit.niso.org/?s=contributor+roles
74. PROV-O: The PROV Ontology. Accessed July 30, 2020. https://www.w3.org/TR/prov-o/
75. FAIRsharing. Accessed July 30, 2020. https://fairsharing.org/standards/
76. Disciplinary Metadata | DCC. Accessed July 30, 2020. https://www.dcc.ac.uk/guidance/standards/metadata
77. RDA Metadata directory. Accessed July 30, 2020. http://rd-alliance.github.io/metadata-directory/
78. CEDAR Better metadata means better science - metadatacenter. Metadata Center. Accessed July 30, 2020. https://metadatacenter.org/
79. Component Metadata | CLARIN ERIC. Accessed July 30, 2020. https://www.clarin.eu/content/component-metadata
80. Data Catalog Vocabulary (DCAT) Namespace. Accessed July 30, 2020. https://www.w3.org/ns/dcat
81. Welcome to the Data Documentation Initiative | Data Documentation Initiative. Accessed July 30, 2020.

FAIRsFAIR “Fostering FAIR Data Practices In Europe” has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558

https://ddialliance.org/
82. Metadata for Description, Discovery & Contextualisation. Agricultural Information Management Standards Portal (AIMS). http://aims.fao.org/activity/blog/metadata-description-discovery-contextualisation-check-rda-metadata-catalog
83. DCMI: Application Profile. Accessed August 3, 2020. https://www.dublincore.org/resources/glossary/application_profile/

FAIRsFAIR "Fostering FAIR Data Practices In Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 grant agreement 831558