KATSAUS

Supporting FAIR data: categorization of research data as a tool in data management

Jessica Parland-von Essen

University of Helsinki

parland@csc.fi

https://orcid.org/0000-0003-4460-3906

Katja Fält

Tampere University of Technology

katja.falt@tut.fi

https://orcid.org/0000-0002-6172-5377

Zubair Maalick

CSC – IT Center for Science

https://orcid.org/0000-0002-0975-1471

Miika Alonen

Aalto University

https://orcid.org/0000-0002-0065-0017

Eduardo Gonzalez

CSC – IT Center for Science

https://orcid.org/0000-0003-1400-0995

The demand for implementation of the FAIR data principles is in many cases difficult for a researcher to adhere to in efficient ways due to lacking tools. We suggest categorizing data in a more extensive and systematic way with focus on the inherent properties of the data as means to enhance research data services. After discussing different approaches to categorizing data, we propose a tripartite research data categorization based on the inherent aspect of stability. The three research data types are operational data, generic research data and research data publications. Generic research data is validated data and can be cumulative, i.e. data can be added without versioning, however if it is dynamic it should be versioned. Generic research data should be separated from immutable dataset publications that are published for reasons of reproducibility of specific research results.

Keywords: research data management; FAIR principles; persistent identifiers; data citation; open science

Persistent identifier: https://doi.org/10.23978/inf.77419

The importance of digital research data as an essential research output is growing due to more strict requirements on reproducibility of results and reusability of data. There are different ways of publishing research data, in journals, data journals and repositories, depending on disciplinary tradition and requirements of research funders and research data policies. In the end they all should serve the need for reproducibility by enabling sustainable data citation, which should be a strict requirement for all services and data publication within research.

The growing amount of data requires professional data management infrastructure and services that facilitate proper use and sharing of data. Digital data, discussed in this article, is digitally generated or stored via digital means. It has specific features, that arise from its impermanent and unstable character – digital data can change over time or disappear altogether, especially if not properly curated and assigned a long-lasting reference, i. e. persistent identifier (or PID, see figure 1). Typically a persistent identifier has two components: a unique identifier; and a service that locates the resource over time even when it's location changes. The first helps to ensure the provenance of a digital resource (that it is what it stands for), whilst the second, a resolver, will ensure that the identifier resolves to the correct current location. The point is, that the dataset the persistent identifiers represent should always be exactly the same. When it comes to data, the persistent identifier should resolve to a landing page with metadata about the resource ("DOI handbook," n.d.).

The amount of digital data is increasing rapidly, and academia is already facing a "data deluge" when it comes to research data like measurement or simulation data that is often produced in vast quantities. Many repositories that hold and curate data today maintain copies of data files that are immutable. Especially services that are domain-agnostic struggle with creating systems that really serve diverse purposes within the realm of research and its ever growing and dynamic data. The more generic the service, the more difficult to reach documentation with enough detail to enable reuse of data.

There are two main objectives of research data management that are discussed in this article. First, there is a need for organising data management practices in and between research organisations and infrastructures to cover and bridge different needs and types of data. This is the data managers perspective. Second, researchers and creators of datasets need to be properly cited for attribution but also for enabling the reuse of data and replication. This is the perspective of the researcher, and its most clear manifestation is data citation. Hence, a reliable system of citation should be developed and implemented as part of research

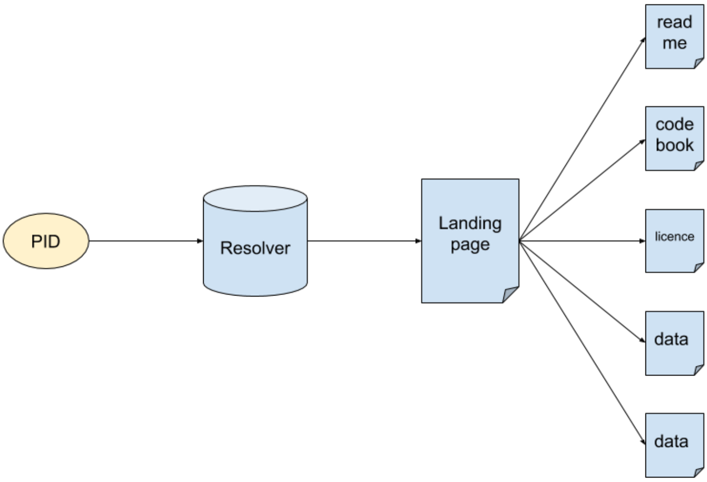


Figure 1: Two-tier persistent identifier. The research dataset as publication in a traditional sense. The identifier points to a specific immutable dataset, enabling data citation and validation of the research.

data management infrastructure. The existing guidelines for citation such as the FAIR data principles (Wilkinson et al., 2016, also see Figure 2) and the Research Data Alliance (RDA) guidance for citing dynamic data both aim at pushing the research culture towards data that is findable, accessible, interoperable and reusable. The growing demand for implementation of the FAIR data principles gives us a great challenge to fix data citation, because they include very explicit requirements regarding extensive management and use of persistent identifiers ("Force11," 2014; Laine & Nykyri, 2018; Mons, 2018). The aim of this article is to identify ways for data management to provide efficient citation technologies and match the FAIR principles.

Demand for better research data management today also stems from research funders. During the last years funders have expressed a nuanced and conscious approach towards Open Science and Open Data. In the European Union the policy follows the principle ”as open as possible, as closed as necessary” and focuses on encouraging sound data management as an essential part of research best practice (H2020 Online Handbook). Openness is seen as one component of FAIR data (Spichtinger, 2016). This can also be considered a reaction to the discussions about replication crisis within the academic community, as well as being well in line with the European Commission's priority of creating a Digital Single Market by unlocking online opportunities and the renewal of the PSI directive (Baker, 2016; European Commission, n.d.). The European Research council demands, that projects must present a data management plan that address following issues (European Research Council (ERC), 2017):

1. Making data Findable,
2. Making data openly Accessible,
3. Making data Interoperable,
4. Increase data Re-use,
5. Allocation of resources and data security.

The Australian National Data Service ANDS has described the FAIR principles as useful because they

- support knowledge discovery and innovation
- support data and knowledge integration
- promote sharing and reuse of data
- are discipline independent and allow for differences in disciplines
- move beyond high level guidance, containing detailed advice on activities that can be undertaken to make data more FAIR
- help data and metadata to be 'machine readable', supporting new discoveries through the harvest and analysis of multiple datasets (ANDS).

Persistent identifiers, thorough documentation and machine actionable linking and citations are at heart of the principles. Still, for instance the RDA guidance to citing dynamic data is in many cases in practice impossible for a researcher to adhere to, if a suitable domain specific infrastructure is not available. It might not be technically or contractually possible do download and archive datasets or allocate persistent identifiers to database queries. We argue, that by analysing data categorization, it is possible to find ways of organizing generic research data management (RDM) services and infrastructures in an adequate way on both a local level and at a larger scale and offer better support for data citation.

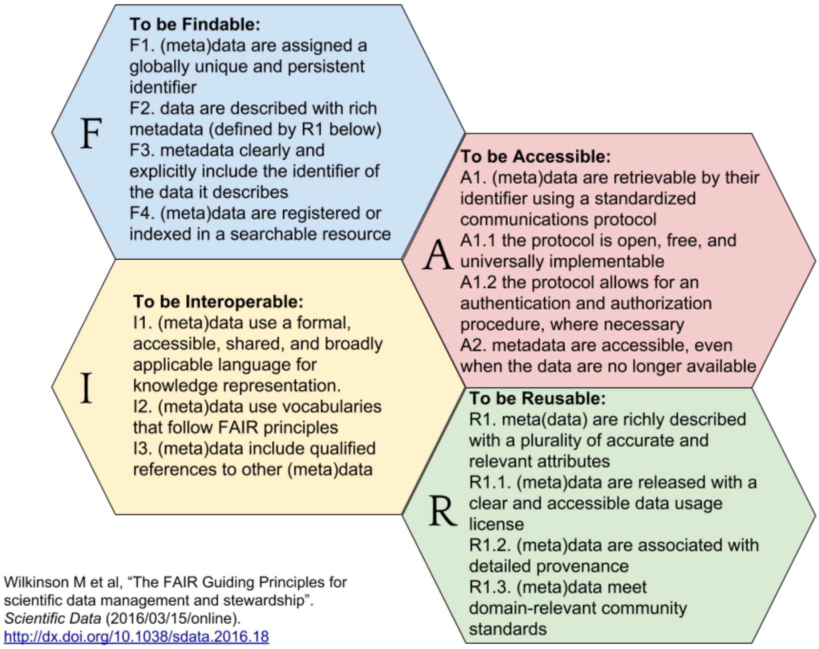
Both data and requirements are diverse. We therefore start by looking at different ways of categorizing research data: from a traditional technical point of view, from a contextual point of view and by looking at the inherent characteristics of the data. We also in each case briefly discuss relevance and comment on how these approaches are used in some of the most common relevant metadata formats. Each of the three ways of categorizing data will be presented and discussed in its own chapter.

We think it would make it easier to plan and manage heterogeneous data resources within organizations using categorization as a tool for structuring data. We find it would be purposeful to focus on the inherent traits of the data, since these are resilient and still underrepresented in the most common metadata formats. We delve a bit deeper into this way of data categorization in the chapter about the three research data types we present as an, as we find, useful way of implementing categorization. This new tripartite research data categorization is based on the stability of the data. The categories are operational data, generic research data and research data publications. Finally, we turn back to our ultimate test and use case, namely the FAIR data principles and the researchers need for trustworthy and sustainable citation. We discuss especially citing dynamic data which poses the most challenges to traditional archival and library thinking.

In the final conclusions we make some reflections on what data categorization could mean for the producers of the RDM services. By addressing conflicts between the data deluge in research data and the traditional static, archival way of looking at data (where a new version always constitutes a complete new copy of a dataset), we hope to achieve more appropriate ways to handle the need for trustworthy but efficient data management and at the same time meet the researchers' need of citation and scientific reproducibility.

Technical aspects: data categorization according to modality or format

The most obvious and traditional way of doing data categorization is perhaps by looking at the technical dimensions of the data. Research data can come in different modalities for humans to consume via ears or eyes and with different sets of natural languages or coding that we can interpret with our senses. In the DOI standard, used with the common persistent identifier, this is called *mode* ("DOI handbook," n.d.). But this categorization is only in part relevant when working with computational research methods. Even cultural studies in digital humanities are today done with algorithms as well as with heuristic methods. Media convergence is in fact one of the main aspects in computer aided research. One technical file format can usually be converted or migrated to another, even across modalities. For instance, linguistic studies have been made by literally picturing sound, and other kinds of visualisations also easily



To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Wilkinson M et al, "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data* (2016/03/15/online). http://dx.doi.org/10.1038/sdata.2016.18

Figure 2: The FAIR data principles.

transgress modalities, similarly to when numeric data is presented as audio. Machine learning transgresses modalites and brings forth new insight precisely thanks to that. (Manovich, 2001; Digens, 2014) Modality and technical format are relevant but not sufficient ways to group resources into categories.

The technical labeling and classification of data is still basic and has perhaps the longest history. Dublin Core, in the 1990's, was the first generic metadata format created for a digital environment and considered as a core set of elements for digital data. Since Dublin Core was strongly influenced by library standards, sensory modality was given a prominent role and consequently the *format* is based on IANA media type terms ("IANA media types," n.d.). The *type* classification on the other hand is impaired by a certain ontological confusion, mixing collections, events, interactive resources and datasets into the same list of controlled values (Dublin Core Metadata Initiative, 2012). Dublin Core elements or today the complementary DCMI terms are often included in other metadata schemas. For instance, the DataCite metadata schema, created for research data, contains an open ResourceType element, which has a general type list similar to the one in the DCMI terms (DataCite Metadata Working Group, 2017).

The Dublin Core way of categorizing data is well established but not necessarily always sufficient or expedient for research. Dublin Core does not require immutable and reusable resources needed to replicate a certain research operation. The format is, on the contrary, flexible enough to work in the unstable digital environment, well embracing such fuzzy objects as services and interactive resources, of which life cycle and technical provenance information as well as documentation of versioning is much up to what is appropriate for each context or domain. However, for research this is not enough, as there is a need to be able to handle different types of static and dynamic data. Data can instead be domesticated for scientific use. For research there is a need to distinguish between growing datasets that are created and collected in systematic ways and the immutable, frozen data publication that underpins a specific research outcome.

Contextual aspects: data categorization according to origin or quality

Some ways of creating data categories are clearly contextual and these should be identified and kept apart from the purely technical or otherwise inherent traits of the data. If we look at a research context, data can for instance be categorized on the basis of how it has been created or collected. From this perspective, data can be divided into four such categories: observational, experimental, simulation and derived compiled data ("Research guides," 2017). Sometimes a fifth category, reference or canonical data, is included (MANTRA, 2017, Data Types & File Formats), Observational data is usually captured in real time and cannot be reproduced or recaptured. Examples of observational data are sensor readings, survey results, telemetry or human observation. Experimental data is typically collected under controlled conditions in a laboratory. This type of data includes gene sequences, microscopy and spectroscopy. Simulation data includes data generated by imitating the operation of a real-world process or system using mathematical models. Such data is usually generated for climate and economic models, chemical reactions or seismic activity. Simulation data is often voluminous and relatively easy to reproduce, provided the process is well documented. Derived compiled data involves using existing data points from different data sources to create data through transformation. This type of data is usually reproducible but often very expensive and time-consuming. Derived or compiled data includes, for example, 3D models, compiled databases or information deriving from text mining. The fifth data type, reference or canonical data is usually peer-reviewed and often published and/or curated collection of datasets. Examples of reference/canonical data include gene sequence databanks, chemical structures or census data.

In Finland, the national research data services fairdata.fi have defined a controlled vocabulary for classifying data according to *use type*, which defines a context according to a specific use category. This was called for when implementing the PREMIS preservation metadata standard in the Finnish long-term preservation services ("National digital preservation services," n.d.; "PREMIS: Preservation Metadata Maintenance Activity (Library of Congress)," n.d.). The Finnish understanding of a research dataset implies that a research dataset *by definition* contains several parts or objects (i.e. files) and that the same file or object can be a part of several datasets. To avoid unnecessary redundancy and to guide the user, the function of each file is specified in the metadata of each dataset when created, which constitutes a use context for the object. The values are *source, output, configuration, publication, method,* and *rights.* By using a controlled vocabulary, the different parts of the dataset are both identifiable by humans and machine actionable. It is then this entity of the complete dataset, that is allocated a two-tiered persistent identifier that can be used for citation ("Metax research datasets," n.d.).

Another contextual aspect is related to the concepts of primary and secondary data. These are not unambiguous exactly because they are in fact contextual categories. A common distinction within social sciences is to divide data into primary and secondary data based on the mode of collection. Primary data is typically collected first hand by the researcher for the purpose of research (for example, survey data, experimental data). Secondary data is usually collected by someone else for purposes other than research (such as census data). But there are differences within academic disciplines in the ways these data types are understood even within in the humanities and social sciences. Whereas a historian easily interprets a primary dataset as stemming from an activity or governmental operations and the secondary data as derivative of the primary data, a social scientist might categorize primary data as a dataset created by herself and a secondary, not as by definition only created by someone else, but also as a validated generic dataset created using a standardised instrument or operating procedure (Hox & Boeije, 2005; MANTRA, 2017). Where the historian tries to look at life cycle of the source and the amount of layers of interpretation, the social scientist might focus on the distinction self-made versus reuse of a dataset.

The reason data was originally created is an important part of its provenance information, but it does not *a priori* limit or define how it can be used. For instance, metadata in itself is often considered to be valuable research data and transformed from operational data to generic research data and maybe even a research dataset publication. The above mentioned Finnish *use type* describes this quality of a dataset that actually is *not* an inherent property of a dataset, but tied to a certain context and originates from the function of the dataset in a specific, given research process. This complicates clear-cut categorizations based on the properties of the data. If the categorization of research data into primary and secondary data is considered related to the context of the dataset, either how it was created or how it is used, it is not a property of the data itself, but always tied to the research question ("Fairdata.fi," n.d.; "Metax research datasets," n.d.; "National digital preservation services," n.d.).

There is a similar problem with the categorization of datasets into qualitative and quantitative datasets, while these terms still state something relevant about how the data is structured. Quantitative data is usually numerical and expresses a certain quantity, amount or range. Qualitative data, on the other hand, is non-numerical and describes the uniqueness the object possesses. ("FIU libraries," n.d.) Still there is an inevitable contextual element in this categorization, similar to the discussion on format. A "qualitative" dataset can be mined and a "quantitative" dataset can be subjected to a heuristic analysis or close reading.

Raw data is a also a term in common use that has a strong context dependence. Within a specific research area, it might be clear amongst colleagues what is meant, but it could be a mistake to simply understand raw data as input or source data in any research process. Often it means data that is not manipulated, validated or cleaned, but the scope might vary considerably between research fields and data types. Raw data can also refer to primary data which means data collected directly from the source such as instrument readings, experiments or survey replies (Mclver, 2011). Once primary data is altered, it becomes secondary data. Altered versions are often described as levels (ESA, n.d.; "Pericles," n.d.). Still, this altered set of primary data may be considered raw data for another stage of the research. This ties into the discussion above about primary and secondary data and proves how confusing the terminology can be.

Active data is not directly related to raw data, but usually means data that is immediately and locally available for an application software and can be used without any modification or reconstruction (Business Dictionary, n.d.; "What is active data?" n.d.). There have also been other efforts to describe the qualities or aspects of data, not in an evaluative way but considering the properties of the data. A growing list of V's for different aspects of data, created for big data, beginning with the three (velocity, volume and variety) and culminating in 42 V's shows how diverse active data can be (Shafer, 2017). The "activeness" of data is actually more related to the next type of aspects we will discuss in the next chapter, namely the inherent characteristics of data.

Inherent aspects: data categorization according to availability, stability and semantic structure

There are some properties of data in a research context, that are neither questions of technical format or coding, nor rising from the context. These are inherent qualities of the data, like whether the data is considered sensitive and therefore cannot be openly available. These are not questions that are decided on scientific grounds, rather the assessment can be ethical or legal. The context is not a research context, but tied to cultural and societal values and arguments, that result in a certain categorization of the data and its availability and its has profound bearing on the data management. Because the inherent properties are stable and often have basic technical consequences, they are a good way of approaching research data categorization from a research data management point of view. They can be derived just by looking at the data itself or relevant legal documentation. The sensitivity is the ethically most important aspect.

As mentioned, one way to categorize research data is according to sensitivity or openness. This is important, because it has immediate and long reaching consequences for the information systems. Data can be completely open (or available) or its use can be restricted due to different legal, contractual or ethical reasons. For this, there are several metadata elements available, but they tend to mix technical access with licenses and rights statements. To enable better machine actionability, Finnish metadata for rights aim at making a distinction between grounds for restriction, licenses (terms of use) and access rights (ATT, 2017). Openness is a trait of the data that is neither context dependent nor technical.

Secondly, there is the aspect of stability, whether the data is dynamic or static and what happens if it is modified. This aspect is of special interest for citation and reproducibility. It will be discussed more closely in the following section of this article, since it is an important element in how research data management should be developed.

Thirdly, things like sampling frequency, grouping, or classification can also be considered inherent features that are not simply technical, functional aspects of the data, but relate to semantics and the quality of the data.

As shown in the arrangement of all the different approaches in table 1, the common metadata formats do not support expressing the inherent traits of data very well compared to their diversity and relevance for the management of the data.

	Trait	Dimension	Example
Technical	language, coding	human interpretation	DCMI Language
	format, MIME type	software related	DCMI Format, a RDF datatype
	modality	video, audio, text etc.	DCMI Type
Contextual	origin	observational, experimental, simulation, derived compiled data	
	use category	source, output, method, metadata	ATT use category
	provenance, lifecycle	primary, secondary, data level, qualitative, quantitative	DDI DataType
Inherent	access type	open data, sensitive data etc.	DCMI Access rights
	research data type	generic data, generic research data, research data publications	
	semantic structure	coherence, levels of measurement, classifications, groupings	

Table 1: Summary of the ways of categorizing data. Dimensions of the characteristics of data in research with examples on metadata.

The three research data types

We will now look a bit closer at the second type of inherent properties, namely the stability of the data, which is the most challenging for creating FAIR data and supporting trustworthy data citation. A categorization based on whether it is active, valid generic research data or a dataset publication, also brings some strength to lifecycle modelling research data. In table 2 the three types are explained with some examples.

	Operational data	Generic research data	Research dataset
Description	Data for any use, private or government owned, might fall within PSI.	Produced by/with/for researchers, validated, good quality, well documented, might be raw or processed.	Dataset produced for a certain research question. Might be highly processed, reuse difficult unless mature field. Main purpose in assesment and reproducibility.
Format	May be dynamic mature solutions, active or even hot data.	Coherent and well documented formats. Data should be quite stable with versioning. Should be possible to cite and enable reproducible research.	Usually in files, but might also be a database with applications. Citation doesn not require date. Two tier resolver for identifier and landing page with metadata available even after data is gone. Might have defined lifespan.
Examples	weather data; data catalogue; big data from social media	corpora; time series of experimental or observational data from technical instruments; similar social or clinical surveys	data paper; data cited in article and published in Zenodo, EUDAT B2Share, other or journal repository

Table 2: Three types of data in a research context

All of these data types might well exist in the same research organization

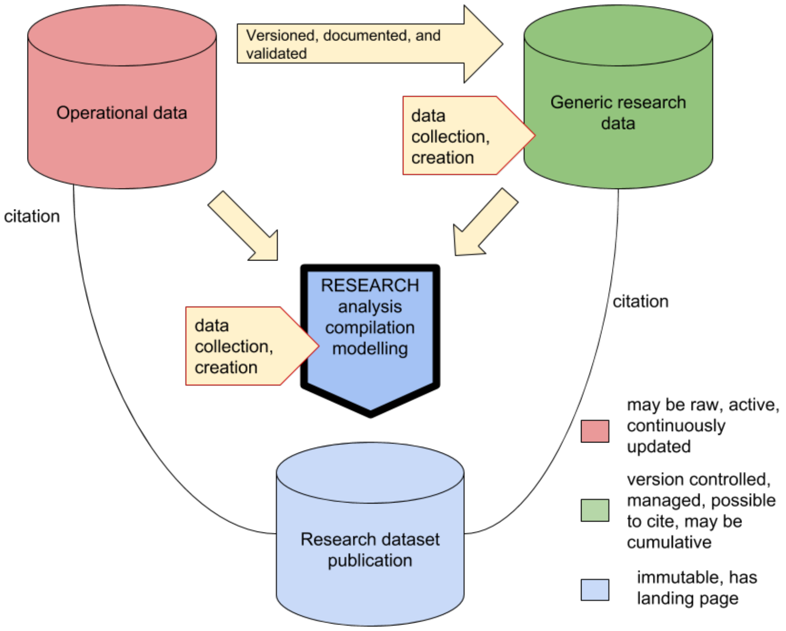

Operational data
Versioned, documented, and validated
Generic research data
data collection, creation
citation
RESEARCH analysis compilation modelling
data collection, creation
citation
Research dataset publication
🟥 may be raw, active, continuously updated
🟩 version controlled, managed, possible to cite, may be cumulative
🟦 immutable, has landing page

Figure 3: The three categories of data in the research process.

and they often do. Firstly, many research organizations and cultural heritage institutions manage data that is in fact operational and they might be obliged to publish it as open data, like their catalogs. In figure 3, it is represented as red. In the end, this data can become traditional, immutable research datasets, that are straightforward to publish and cite with a normal two-tier persistent identifier as the output of a scientific study (blue in figure 3). If this kind of dataset is modified, a new persistent identifier is allocated, which is then linked as a new version. If the data is retracted, a tombstone page will be created.

But there is also a third type of data relevant for research. These datasets might change in well managed and documented ways as cumulative datasets that evolve in a very specific way (green in figure 3). They consist of validated and well documented data for research and change only in a given manner, i.e. by growing. Previous versions of data may not be retracted, and if it is, it has to be very well justified because it might endanger the research. Generic research data is safe to cite. In practice, this could be achieved by carefully controlled changes, for instance by only allowing additions of files. This is a simple way to ensure that previous data is unchanged. These datasets can be ”open” until they are ”closed”, when they become completely immutable and conform to normal versioning practice. This data type has been very well described by the Committee on Earth Observation Satellites (CEOS Data Stewardship Interest Group, 2017). Citation should be advised in a slightly different way for open cumulative datasets. There are many users who could take advantage of this kind of dataset when working with longer time series or campaign workflows while creating and publishing research data. It seems pragmatic to take into account this type of datasets when creating data management services for research (UNIFI, 2018). The dataset can also be versioned or documented in other meticulous ways, so that data can always be recreated.

It might clarify data management considerably if the resources would be classified according to this partitioning. If generic cumulative research datasets are managed in the same catalogues and made available through the same services as individual research dataset publications, it could be a good idea to tag them both for data management reasons (different criteria for creation of persistent identifiers) and for the citation guideline (where citation date is needed).

FAIR data in use: the researchers' point of view

The usefulness of the categorization presented above becomes visible when we take a look at the researchers' needs. Citation is an essential ethical principle in the academic research process further underlined by the demand for FAIR data. The idea behind citing research data is similar to citing articles and books; it ensures that the creators of a dataset get proper attribution and credit. But data citation also enables reproducibility of findings and supports the reuse of data. An effective data citing system enables datasets to be integrated into the scholarly communication, to be properly used, found and managed (Weller, 2011, p. 45). This means that either the dataset has to be preserved as is or it has to be possible to recreate it when needed. This can be challenging if the dataset is subject to frequent updates. Versioning is not always feasible if, for instance, the dataset is an active database rather than an archived file, especially if it is a database that is not primarily created for research use (red rather than green in figure 3).

The national recommendation for data citation was published in Finland in 2018 and it follows international guidelines such as FORCE11. It states that a data reference should consist of the following elements: creator, title, host organisation, publication time and/or date, and persistent identifier. Useful additional elements are version, resource type, license status, ORCID, and embargo information (Finnish Committee for Research Data, 2018). In other words, it still follows the traditional dataset publication (blue) logic.

DataCite offers a research domain metadata schema that is primarily developed for research purposes and is therefore tightly linked to DOI. This handle based persistent identifier has a strong brand, which has made it easier to deploy and to be used by researchers and research communities. The latter aspect is important, since awareness and skills in research data management vary among end users within the scientific community. For wide adoption and good citation practices branding is therefore important. Users need to be able to easily identify a persistent identifier. DataCite as metadata format will hardly substitute domain specific formats, but it has the capacity to function as a generic fallback and an exchange format. It is mostly developed with dataset publication and citation in mind. DataCite gives the following guidelines (DataCite Metadata Working Group, 2017, p. 12) or options for citing dynamic data:

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

It is also noted that the three first alternatives require unique and persistent identifiers, but we find it important to underline that these cannot necessarily be created by the user without creating a copy by downloading the data and then archiving it somewhere else. All in all, citing dynamic datasets puts enormous demands on the data management infrastructure, as does the RDA recommendation for citing dynamic data (Rauber, van Uytvanck, Asmi, & Pröll, 2016). Still, if we consider that there is a certain type of cumulative dataset where there is only one type of modification, which is addition of uniform data, this would make a distinct special case of case c) that would not endanger the uniqueness of the dataset and reproducibility of research. It would also prevent rampant minting and allocation of two-tier persistent identifiers (ANDS, n.d.) which constitute an unnecessary, underestimated cost and deadweight to manage, if never even used for citation. It seems sensible that persistent identifiers for dynamic data should be created on demand simply because these need extensive metadata to be valid research data. Reproducibility and identification of the dynamic research data can also be achieved with the implementation of query repository that persistently identifies and verifies the query results (Figure 4) as

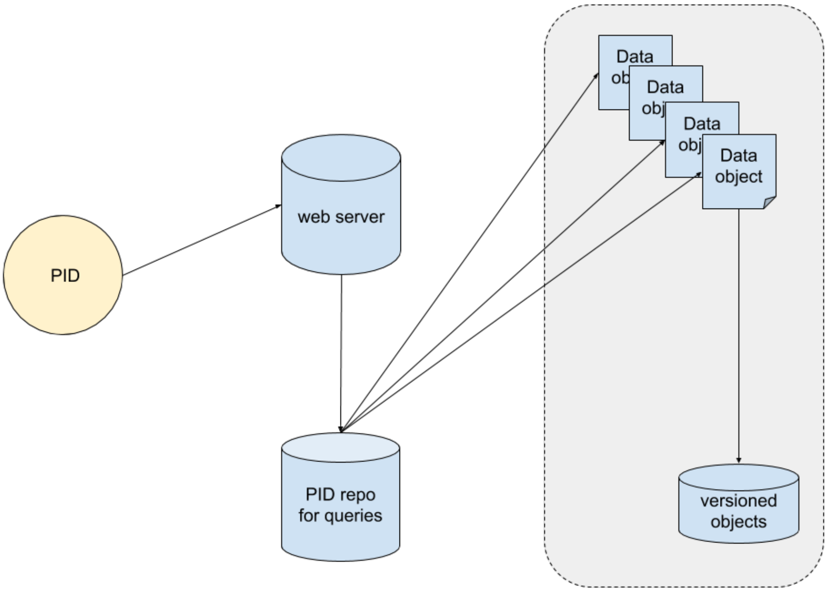


Figure 4: Data citation for dynamic data, case d) above.

described by the RDA working group on Data Citation. There are many ways to solve questions of reproducibility of science. For instance in the Language Bank of Finland a solution with a so called stop-over pages has been introduced to accommodate the needs of data management and citation (Matthiesen & Dieckmann, 2018).

Life cycle events can rarely be automatically documented sufficiently for future scientific use because the research context and questions at large are usually not documented in a structured way. All cases mentioned in the DataCite documentation, except the snapshot, require advanced, trustworthy, high level data management services. These are unfortunately unavailable in most fields of science. This is not only due to immature infrastructures or flaws in research culture, but it is in the very essence of science to be novel and original. If we managed to standardize all knowledge, it would probably prevent or at least constrain future development.

Conclusions

Planning information management in research there are several different aspects or dimensions of data that need to be taken into account. To be able to manage data during its lifecycle and also offer researchers means to produce FAIR data, it is important to support citation and provide clear information about the data.

However, managing different types of data may require different solutions in data curation and preservation. Diligent categorizing of data helps organizing and structuring data management services. Focusing in the inherent traits of the data is both efficient and valid across contexts. Distinguishing different types of datasets and creating tailored, standardised procedures for each type can be done by implementing a categorization between 1) operational and active, 2) generic and dynamic research data and 3) the immutable datasets. The FAIR data principles should be paid attention to planning the services, ensuring that data is findable, accessible, interoperable and reusable. Developing tools and processes for transforming active data to generic research datasets and research dataset publications will give better support for lifecycle management. While repositories today often offer services for frozen data publications, we still need to find solutions for citing dynamic data and managing transitions and references.

Categorizing data more systematically by separating different traits and dimensions would equally be of help when compiling data management guidance and training for researchers. An overview of the here discussed aspects and categories is presented in table 2. It would be easier to instruct researchers if there were standardized, organization-specific procedures in data management and archiving based on categorization since data always comes in many forms. Each data type could, for example, have its own set of procedures for creating metadata, defining the terms of reuse, choosing a suitable archive, citations, and so on. Despite sometimes obvious difficulties in clear-cut categorization and the nebulous character of real-life research data, efforts should be made to clarify the concepts and in implementing them as parts of data management.

We suggest that the contextual traits of data are more clearly separated from its inherent qualities and that generic research data would be separated from datasets that are outputs of specific research focused projects that have to be kept immutable. Managing and citing cumulative datasets should be handled differently from discrete research data publications. Metadata formats and research data architecture should be developed further to better answer the diverse needs in data management and research. Data categorization is an important tool that is currently underutilized in creating infrastructures and services that could enable wider deployment of FAIR data compliant practices.

References

ANDS. (n.d.). Guides and resources. Persistent identifiers. *Australian national data service*. https://www.ands.org.au/guides/persistent-identifiers-expert

ATT. (2017). *Oikeuksien hallintaan liittyvät metatiedot -selvitys*. Opetus- ja kulttuuriministeriö. http://urn.fi/URN:NBN:fi-fe201702101528

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452. https://doi.org/10.1038/533452a

Business Dictionary. (n.d.). What is active data? Definition and meaning. *BusinessDictionary.com*. http://www.businessdictionary.com/definition/active-data.html

CEOS Data Stewardship Interest Group. (2017). Persistent identifier best practices. Version 1.2. CEOS/wgiss/dsig/pidbp. http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS/%20Best/%20Practices/CEOS/%20Persistent/%20Identifier/%20Best/%20Practices_v1.2.pdf

DataCite Metadata Working Group. (2017). DataCite metadata schema 4.1. *DataCite Schema*. http://doi.org/10.5438/0014

Digens, A. (2014). These artists are turning space junk into sound art. *Creators*. https://www.vice.com/en_au/article/9anad3/artists-are-transforming-space-junk-into-sound-art

DOI handbook. (n.d.). http://doi.org/10.1000/182

Dublin Core Metadata Initiative. (2012). DCMI terms. *Dublin Core*. http://dublincore.org/documents/dcmi-terms/

ESA. (n.d.). Products and algorithms. *Sentinel*. https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/products-algorithms

European Commission. (n.d.). Digital Single Market. Proposal for a revision of the public sector information (PSI) directive. *Digital Single Market*. https://ec.europa.eu/digital-single-market/en/proposal-revision-public-sector-information-psi-directive

European Research Council (ERC). (2017). *Guidelines on implementation of open access to scientific publications and research data in projects supported by the European Research Council under Horizon 2020. Version 1.1.* http://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf

Fairdata.fi. (n.d.). *Fairdata*. https://www.fairdata.fi/

Finnish Committee for Research Data. (2018). Tracing data: data citation roadmap for Finland. Finnish Committee for Research Data. http://urn.fi/URN:NBN:fi-fe201804106446

Force11: The fair data principles. (2014). *FORCE11*. https://www.force11.org/group/fairgroup/fairprinciples

Hox, J. J., & Boeije, H. R. (2005). Data collection, primary versus secondary. In *Encyclopedia of social measurement* (pp. 593–599). Elsevier. http://hdl.handle.net/1874/23634

IANA media types. (n.d.). *IANA*. https://www.iana.org/assignments/media-types/media-types.xhtml

Laine, H., & Nykyri, S. (2018). Dataviittaamisen tiekartta tutkijalle. *Informaatiotutkimus*, *37*(2). https://doi.org/10.23978/inf.72999

Manovich, L. (2001). *The language of new media*. MIT Press.

MANTRA. (2017). Research data explained. https://doi.org/10.5281/zenodo.1035218

Matthiesen, M., & Dieckmann, U. (2018). Versioning with PIDs. In *CLARIN Annual Conference 2018 in Pisa, Italy.* CLARIN. https://www.clarin.eu/clarin-annual-conference-2018-abstracts

Mclver, J. P. (2011). Raw data. In P. Lavrakas (ed.), *Encyclopedia of survey research methods*. Thousand Oaks: SAGE Publications, Inc. https://doi.org/10.4135/9781412963947.n447

Metax research datasets. (n.d.). *Tietomallit*. https://tietomallit.suomi.fi/model/mrd/CatalogRecord/

Mons, B. (2018). *Data stewardship for open science: Implementing fair principles*. Chapman and Hall/CRC. https://www.crcpress.com/Data-Stewardship-for-Discovery-A-Practical-Guide-for-Data-Experts/Mons/p/book/9780815348184

National digital preservation services. (n.d.). *digitalpreservation.fi*. http://digitalpreservation.fi/en

Pericles. (n.d.). *Pericles*. http://pericles-project.eu/training-module/space-data/space-project-phasing-data-levels-and-data-use/processing-levels/

PREMIS: Preservation Metadata Maintenance Activity (Library of Congress). (n.d.). *PREMIS*. https://www.loc.gov/standards/premis/

Rauber, A., van Uytvanck, D., Asmi, A., & Pröll, S. (2016). *Identification of reproducible subsets for data citation, sharing and re-use*. Research Data Alliance. https://www.rd-alliance.org/system/files/documents/TCDL-RDA-Guidelines_160411.pdf

Research guides: Data module #1: What is research data? (2017). https://libguides.macalester.edu/c.php?g=527786/&p=3608643

Research methods help guide. (n.d.). https://library.fiu.edu/friendly.php?s=researchmethods/datatypes

Shafer, T. (2017). The 42 V's of big data and data science. *Elder Research*. https://www.elderresearch.com/blog/42-v-of-big-data

Spichtinger, D. (2016). Open / fair research data in horizon 2020. https://ec.europa.eu/easme/sites/easme-site/files/open_fair_research_data_in_h2020.pdf

UNIFI. (2018). Avoin tiede ja data. Toimenpideohjelma suomalaiselle tiedeyhteisölle. Universities Finland UNIFI. http://urn.fi/URN:NBN:fi-fe2018052424593

Webopedia "active data". (n.d.). *Webopedia*. https://www.webopedia.com/TERM/A/active_data.html

Weller, M. (2011). *The digital scholar: How technology is transforming scholarly practice*. Bloomsbury Academic. https://doi.org/10.5040/9781849666275

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18