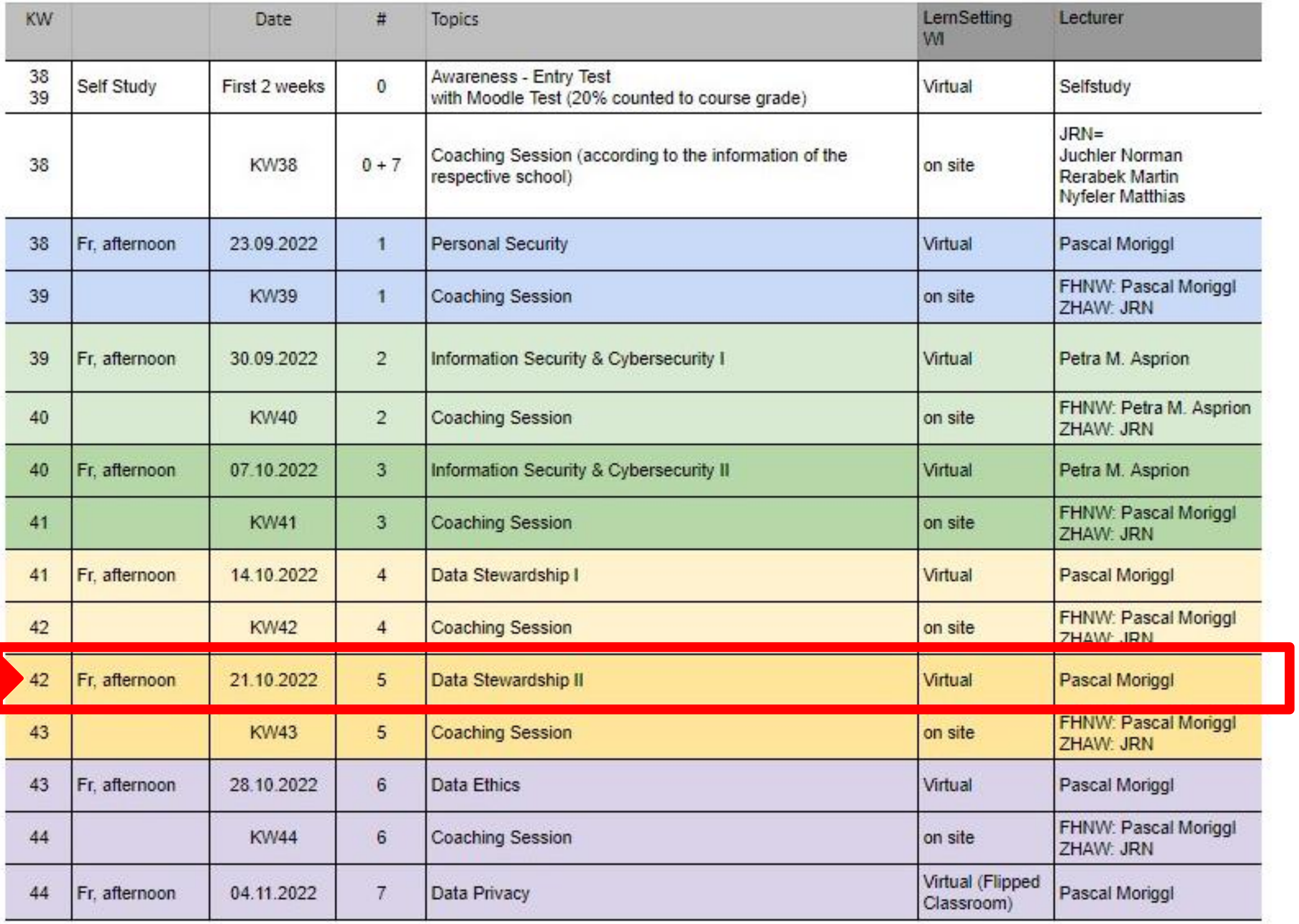
University of Applied Sciences and Arts Northwestern Switzerland
School of Life Sciences

member of swissuniversities

S5 – Data Stewardship II

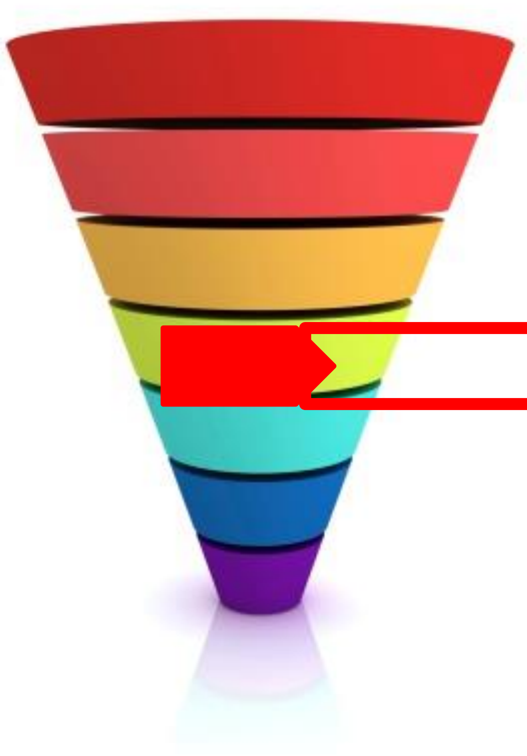
D4 – Data and Ethics

Schedule



KW		Date	#	Topics	LernSetting WI	Lecturer
38 39	Self Study	First 2 weeks	0	Awareness - Entry Test with Moodle Test (20% counted to course grade)	Virtual	Selfstudy
38		KW38	0 + 7	Coaching Session (according to the information of the respective school)	on site	JRN= Juchler Norman Rerabek Martin Nyfeler Matthias
38	Fr, afternoon	23.09.2022	1	Personal Security	Virtual	Pascal Moriggl
39		KW39	1	Coaching Session	on site	FHNW: Pascal Moriggl ZHAW: JRN
39	Fr, afternoon	30.09.2022	2	Information Security & Cybersecurity I	Virtual	Petra M. Asprion
40		KW40	2	Coaching Session	on site	FHNW: Petra M. Asprion ZHAW: JRN
40	Fr, afternoon	07.10.2022	3	Information Security & Cybersecurity II	Virtual	Petra M. Asprion
41		KW41	3	Coaching Session	on site	FHNW: Pascal Moriggl ZHAW: JRN
41	Fr, afternoon	14.10.2022	4	Data Stewardship I	Virtual	Pascal Moriggl
42		KW42	4	Coaching Session	on site	FHNW: Pascal Moriggl ZHAW: JRN
42	Fr, afternoon	21.10.2022	5	Data Stewardship II	Virtual	Pascal Moriggl
43		KW43	5	Coaching Session	on site	FHNW: Pascal Moriggl ZHAW: JRN
43	Fr, afternoon	28.10.2022	6	Data Ethics	Virtual	Pascal Moriggl
44		KW44	6	Coaching Session	on site	FHNW: Pascal Moriggl ZHAW: JRN
44	Fr, afternoon	04.11.2022	7	Data Privacy	Virtual (Flipped Classroom)	Pascal Moriggl

Where are we at? Big Picture



1. Secure myself
2. Secure my Organisation
3. Keep my project clean through data management
4. Keep my project data clean through FAIR
5. Do the right* thing with the data (Ethics)
6. Do the correct thing with the data (Privacy)

Agenda

Part 0 : Repetition Last Week

Part I : Intro FAIR

Part II : Copyright / Licencing

Part III: Data Formatting

Our topic --
Data Stewardship

Data Sharing // Snafu in 3 Short Acts

https://youtu.be/66oNv_DJuPc

University of Applie
School of Life Scien

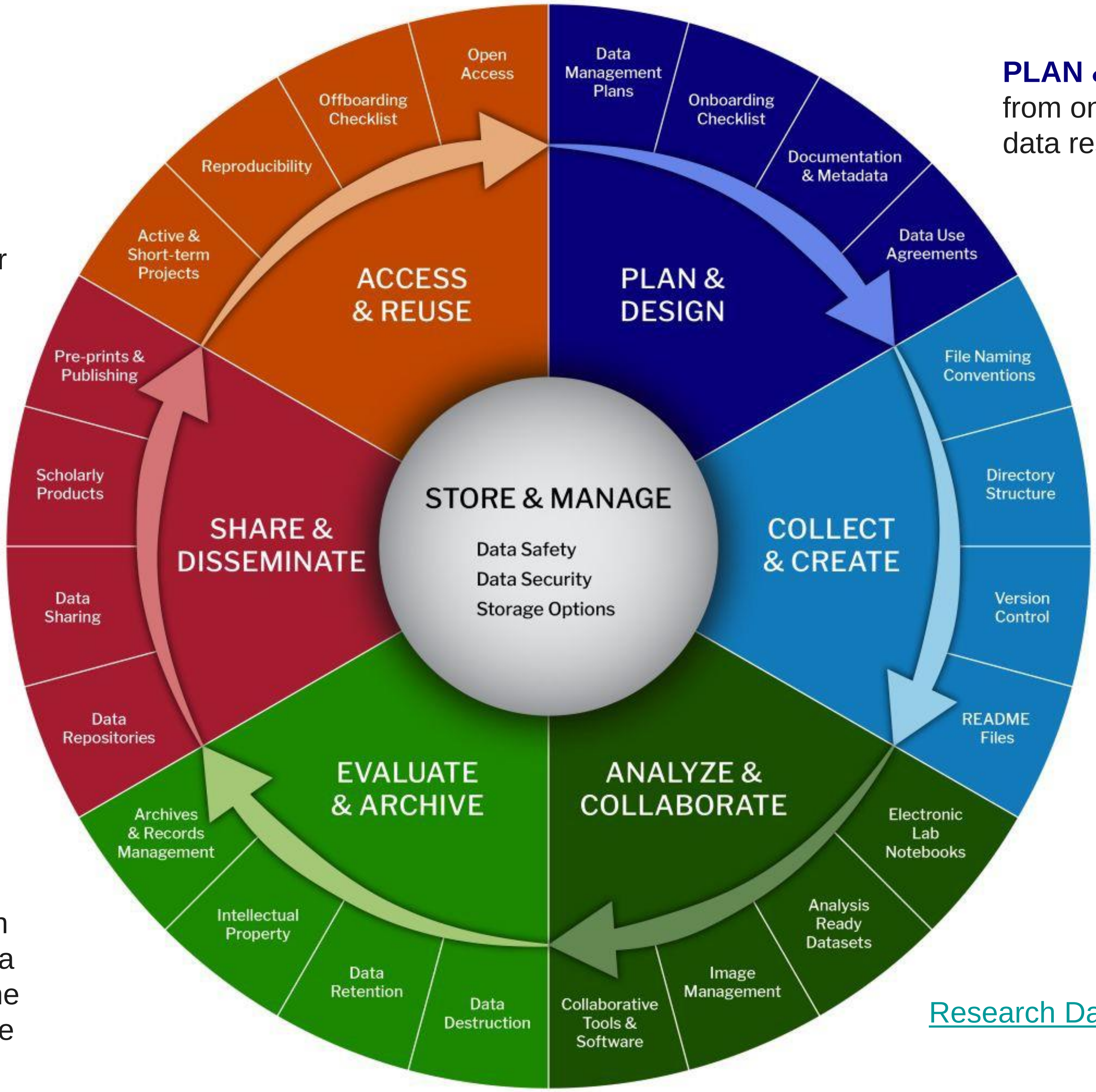
Lecturer 4
Data Stewardship

ACCESS & REUSE: Ensuring the broad utility of your research data efforts for other researchers

SHARE & DISSEMINATE: Establishing and supporting the reach and impact of your data

EVALUATE & ARCHIVE: Identify essential research records and evaluate for retention

STORE & MANAGE: Each stage of the Biomedical Data Lifecycle revolves around the management of data storage



PLAN & DESIGN: Plan processes from onboarding to project closure and data resources

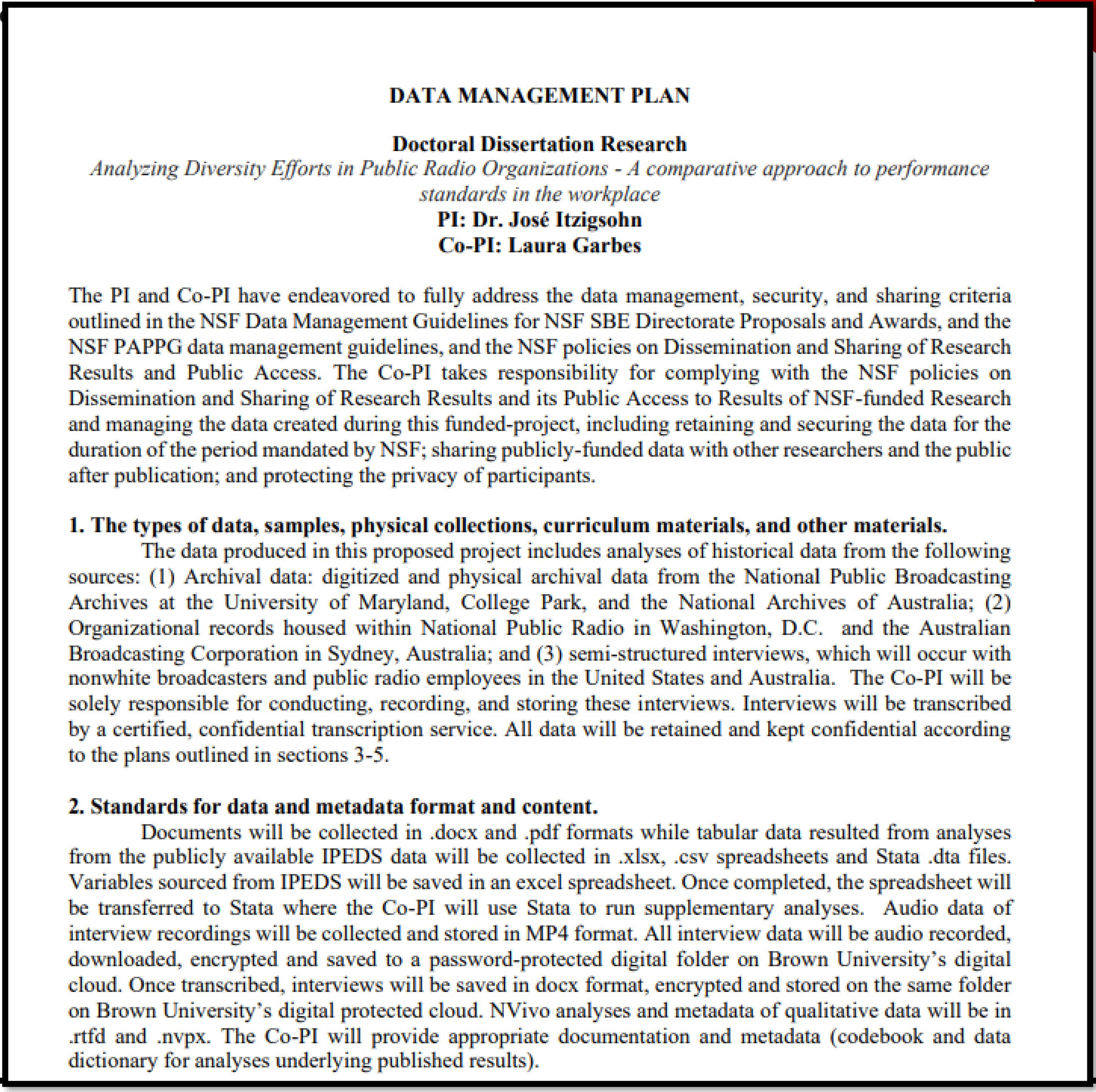
COLLECT & CREATE: Organization and integration of data sets and collection processes

ANALYZE & COLLABORATE: Processing and analyzing data should be collaborative and documented

Research Data Lifecycle by LMA RDMWG

Data Management Plan

A joyful and exciting exercise ☺



DATA MANAGEMENT PLAN

Doctoral Dissertation Research
Analyzing Diversity Efforts in Public Radio Organizations - A comparative approach to performance standards in the workplace
PI: Dr. José Itzigsohn
Co-PI: Laura Garbes

The PI and Co-PI have endeavored to fully address the data management, security, and sharing criteria outlined in the NSF Data Management Guidelines for NSF SBE Directorate Proposals and Awards, and the NSF PAPPG data management guidelines, and the NSF policies on Dissemination and Sharing of Research Results and Public Access. The Co-PI takes responsibility for complying with the NSF policies on Dissemination and Sharing of Research Results and its Public Access to Results of NSF-funded Research and managing the data created during this funded-project, including retaining and securing the data for the duration of the period mandated by NSF; sharing publicly-funded data with other researchers and the public after publication; and protecting the privacy of participants.

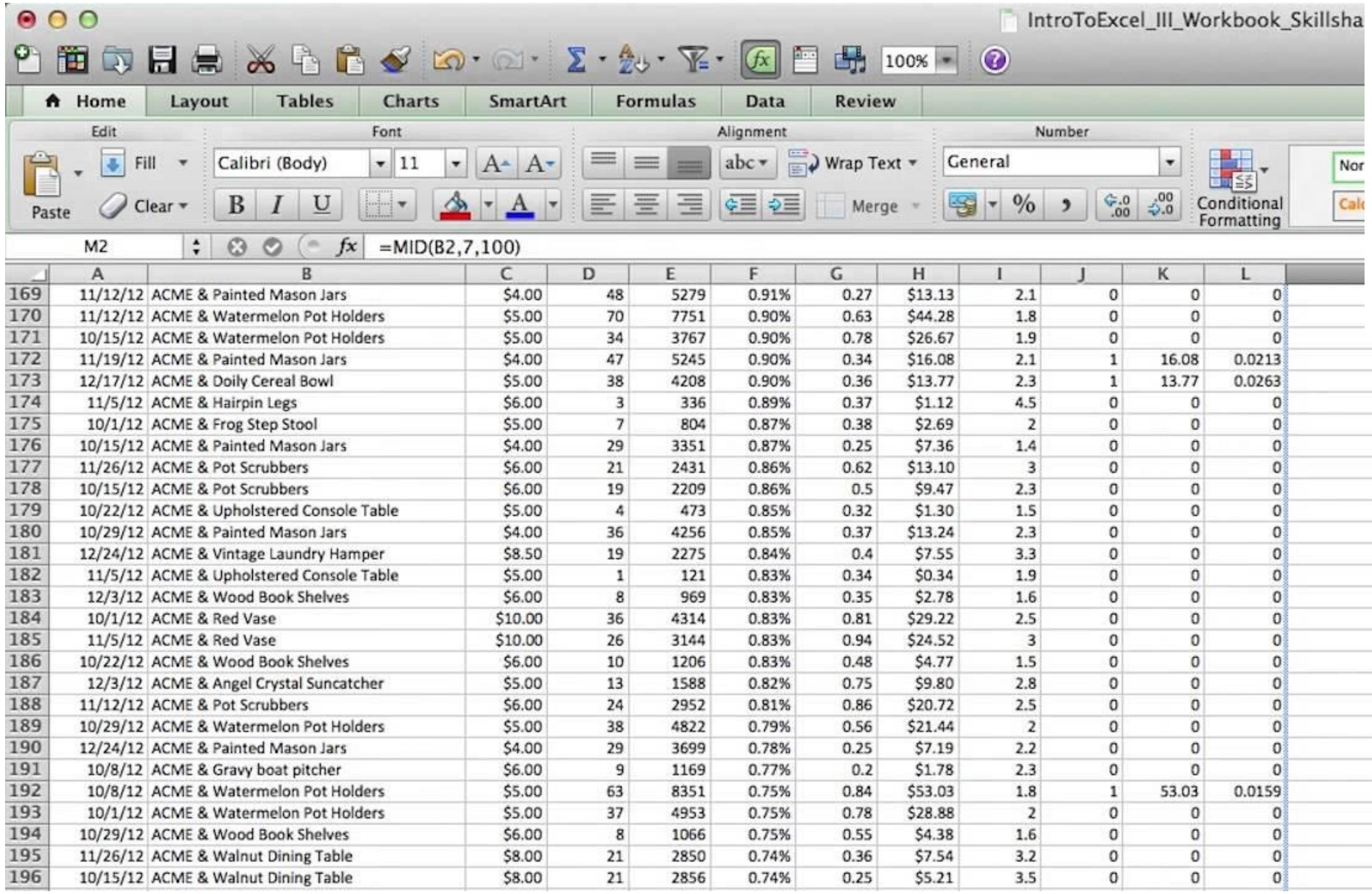
1. The types of data, samples, physical collections, curriculum materials, and other materials.

The data produced in this proposed project includes analyses of historical data from the following sources: (1) Archival data: digitized and physical archival data from the National Public Broadcasting Archives at the University of Maryland, College Park, and the National Archives of Australia; (2) Organizational records housed within National Public Radio in Washington, D.C. and the Australian Broadcasting Corporation in Sydney, Australia; and (3) semi-structured interviews, which will occur with nonwhite broadcasters and public radio employees in the United States and Australia. The Co-PI will be solely responsible for conducting, recording, and storing these interviews. Interviews will be transcribed by a certified, confidential transcription service. All data will be retained and kept confidential according to the plans outlined in sections 3-5.

2. Standards for data and metadata format and content.

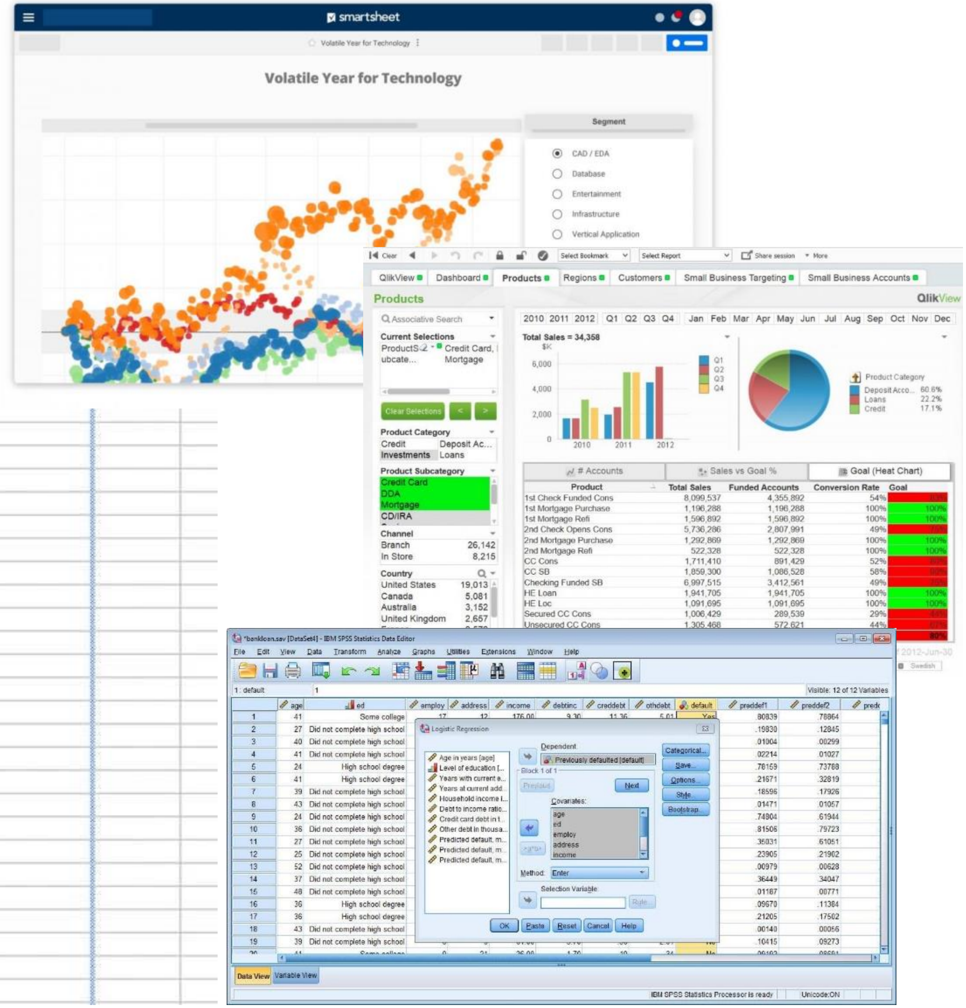
Documents will be collected in .docx and .pdf formats while tabular data resulted from analyses from the publicly available IPEDS data will be collected in .xlsx, .csv spreadsheets and Stata .dta files. Variables sourced from IPEDS will be saved in an excel spreadsheet. Once completed, the spreadsheet will be transferred to Stata where the Co-PI will use Stata to run supplementary analyses. Audio data of interview recordings will be collected and stored in MP4 format. All interview data will be audio recorded, downloaded, encrypted and saved to a password-protected digital folder on Brown University's digital cloud. Once transcribed, interviews will be saved in docx format, encrypted and stored on the same folder on Brown University's digital protected cloud. NVivo analyses and metadata of qualitative data will be in .rtfd and .nvpx. The Co-PI will provide appropriate documentation and metadata (codebook and data dictionary for analyses underlying published results).

Data ~~Management Plan~~



M2 fx =MID(B2,7,100)

	A	B	C	D	E	F	G	H	I	J	K	L
169	11/12/12	ACME & Painted Mason Jars	$4.00	48	5279	0.91%	0.27	$13.13	2.1	0	0	0
170	11/12/12	ACME & Watermelon Pot Holders	$5.00	70	7751	0.90%	0.63	$44.28	1.8	0	0	0
171	10/15/12	ACME & Watermelon Pot Holders	$5.00	34	3767	0.90%	0.78	$26.67	1.9	0	0	0
172	11/19/12	ACME & Painted Mason Jars	$4.00	47	5245	0.90%	0.34	$16.08	2.1	1	16.08	0.0213
173	12/17/12	ACME & Doily Cereal Bowl	$5.00	38	4208	0.90%	0.36	$13.77	2.3	1	13.77	0.0263
174	11/5/12	ACME & Hairpin Legs	$6.00	3	336	0.89%	0.37	$1.12	4.5	0	0	0
175	10/1/12	ACME & Frog Step Stool	$5.00	7	804	0.87%	0.38	$2.69	2	0	0	0
176	10/15/12	ACME & Painted Mason Jars	$4.00	29	3351	0.87%	0.25	$7.36	1.4	0	0	0
177	11/26/12	ACME & Pot Scrubbers	$6.00	21	2431	0.86%	0.62	$13.10	3	0	0	0
178	10/15/12	ACME & Pot Scrubbers	$6.00	19	2209	0.86%	0.5	$9.47	2.3	0	0	0
179	10/22/12	ACME & Upholstered Console Table	$5.00	4	473	0.85%	0.32	$1.30	1.5	0	0	0
180	10/29/12	ACME & Painted Mason Jars	$4.00	36	4256	0.85%	0.37	$13.24	2.3	0	0	0
181	12/24/12	ACME & Vintage Laundry Hamper	$8.50	19	2275	0.84%	0.4	$7.55	3.3	0	0	0
182	11/5/12	ACME & Upholstered Console Table	$5.00	1	121	0.83%	0.34	$0.34	1.9	0	0	0
183	12/3/12	ACME & Wood Book Shelves	$6.00	8	969	0.83%	0.35	$2.78	1.6	0	0	0
184	10/1/12	ACME & Red Vase	$10.00	36	4314	0.83%	0.81	$29.22	2.5	0	0	0
185	11/5/12	ACME & Red Vase	$10.00	26	3144	0.83%	0.94	$24.52	3	0	0	0
186	10/22/12	ACME & Wood Book Shelves	$6.00	10	1206	0.83%	0.48	$4.77	1.5	0	0	0
187	12/3/12	ACME & Angel Crystal Suncatcher	$5.00	13	1588	0.82%	0.75	$9.80	2.8	0	0	0
188	11/12/12	ACME & Pot Scrubbers	$6.00	24	2952	0.81%	0.86	$20.72	2.5	0	0	0
189	10/29/12	ACME & Watermelon Pot Holders	$5.00	38	4822	0.79%	0.56	$21.44	2	0	0	0
190	12/24/12	ACME & Painted Mason Jars	$4.00	29	3699	0.78%	0.25	$7.19	2.2	0	0	0
191	10/8/12	ACME & Gravy boat pitcher	$6.00	9	1169	0.77%	0.2	$1.78	2.3	0	0	0
192	10/8/12	ACME & Watermelon Pot Holders	$5.00	63	8351	0.75%	0.84	$53.03	1.8	1	53.03	0.0159
193	10/1/12	ACME & Watermelon Pot Holders	$5.00	37	4953	0.75%	0.78	$28.88	2	0	0	0
194	10/29/12	ACME & Wood Book Shelves	$6.00	8	1066	0.75%	0.55	$4.38	1.6	0	0	0
195	11/26/12	ACME & Walnut Dining Table	$8.00	21	2850	0.74%	0.36	$7.54	3.2	0	0	0
196	10/15/12	ACME & Walnut Dining Table	$8.00	21	2856	0.74%	0.25	$5.21	3.5	0	0	0



Product Category	
Credit	Deposit Ac...
Investments	Loans

Channel	
Branch	26,142
In Store	8,215

Country	
United States	19,013
Canada	5,081
Australia	3,152
United Kingdom	2,657

Product	Total Sales	Funded Accounts	Conversion Rate	Goal
1st Check Funded Cons	8,099,537	4,355,892	54%	
1st Mortgage Purchase	1,196,288	1,196,288	100%	100%
1st Mortgage Refi	1,596,892	1,596,892	100%	100%
2nd Check Opens Cons	5,736,286	2,807,991	49%	
2nd Mortgage Purchase	1,292,869	1,292,869	100%	100%
2nd Mortgage Refi	522,328	522,328	100%	100%
CC Cons	1,711,410	891,429	52%	
CC SB	1,859,300	1,086,528	58%	
Checking Funded SB	6,997,515	3,412,561	49%	
HE Loan	1,941,705	1,941,705	100%	100%
HE Loc	1,091,695	1,091,695	100%	100%
Secured CC Cons	1,006,429	289,539	29%	
Unsecured CC Cons	1,305,468	572,621	44%	

1 : default 1

	age	ed	preddef1	preddef2
1	41	Some college	.80839	.78864
2	27	Did not complete high school	.19830	.12845
3	40	Did not complete high school	.01004	.00299
4	41	Did not complete high school	.02214	.01027
5	24	High school degree	.78159	.73788
6	41	High school degree	.21671	.32819
7	39	Did not complete high school	.18596	.17926
8	43	Did not complete high school	.01471	.01057
9	24	Did not complete high school	.74804	.61944
10	36	Did not complete high school	.81506	.79723
11	27	Did not complete high school	.35031	.61051
12	25	Did not complete high school	.23905	.21902
13	52	Did not complete high school	.00979	.00628
14	37	Did not complete high school	.36449	.34047
15	48	Did not complete high school	.01187	.00771
16	36	High school degree	.09670	.11384
17	36	High school degree	.21205	.17502
18	43	Did not complete high school	.00140	.00056
19	39	Did not complete high school	.10415	.09273

Learning Goals

- ✓ Increase awareness for FAIR data principles
- ✓ Understand methods and processes to keep your datasets comprehensible
- ✓ Know the key principles of data-level implementation FAIR

What is FAIR?

… three fundamental concepts:

- ✓ the FAIR principles
- ✓ FAIR data
- ✓ FAIRification practices.

What FAIR?

The FAIR principles

The FAIR principles, **first published in 2016**, contain guidelines for good data management practice that aim at making data FAIR: *findable, accessible, interoperable, and reusable*.

"Data" refers in this context to all kinds of digital objects that are produced in research: research data in the strictest sense, code, software, presentations, etc.

What FAIR?

The FAIR principles

Each letter in FAIR refers to a list of principles with a total of 15 principles altogether. **A recent paper on the FAIR principles** introduces the useful distinction between *the four foundational principles* that aim at findability, accessibility, interoperability, and reusability and *the 15 guiding principles* that more explicitly and measurably describe how FAIRness of data can be achieved through technical implementation.

…. Although the FAIR principles originate from the life sciences, they can be applied within all research disciplines…..

The FAIR principles

Key Thoughts:

1) Both humans and machines are intended as digesters of data.

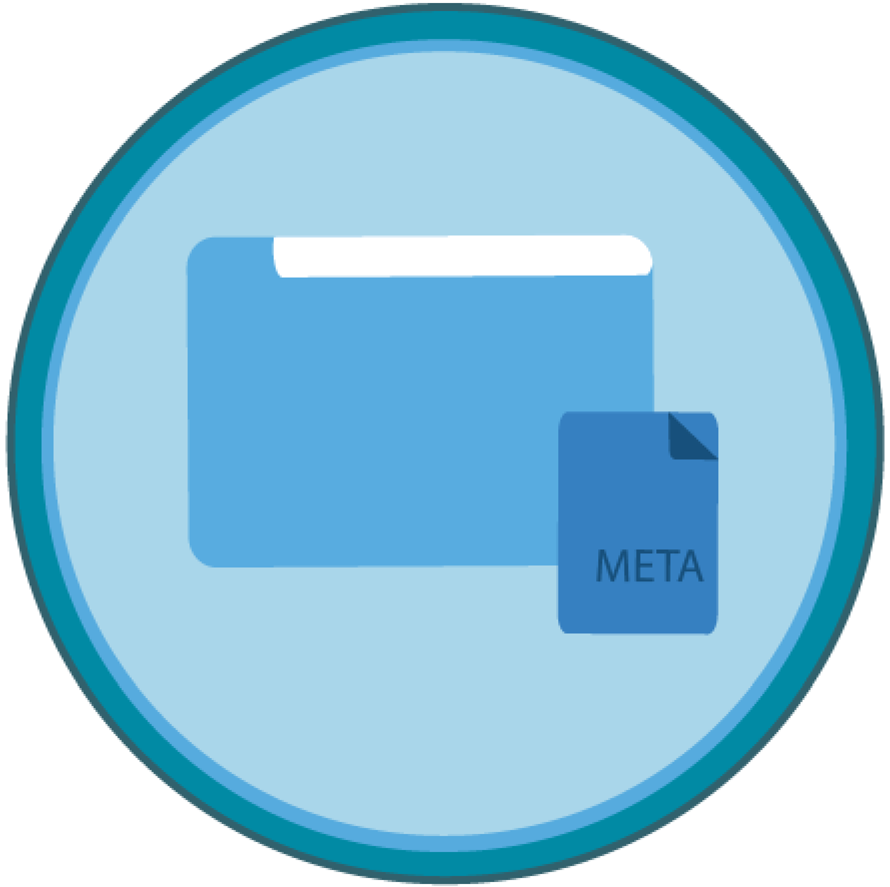
This will lead to the creation of an ecosystem that is fast to respond to change and automatically adapts to new findings or changes: the ***Internet of FAIR Data and Services***. This is the reason for focusing on standards for data, identification mechanisms, data availability, etc.

The FAIR principles

Key Thoughts:

2) The FAIR principles apply to both *data* and *metadata.*

Where metadata are descriptions of or records about data. This is why the term “(meta)data” is stated in the principles.



The FAIR principles

Key Thoughts:

3) The principles are not necessarily about open data.

You can work in a FAIR manner with data that is not intended for public availability.

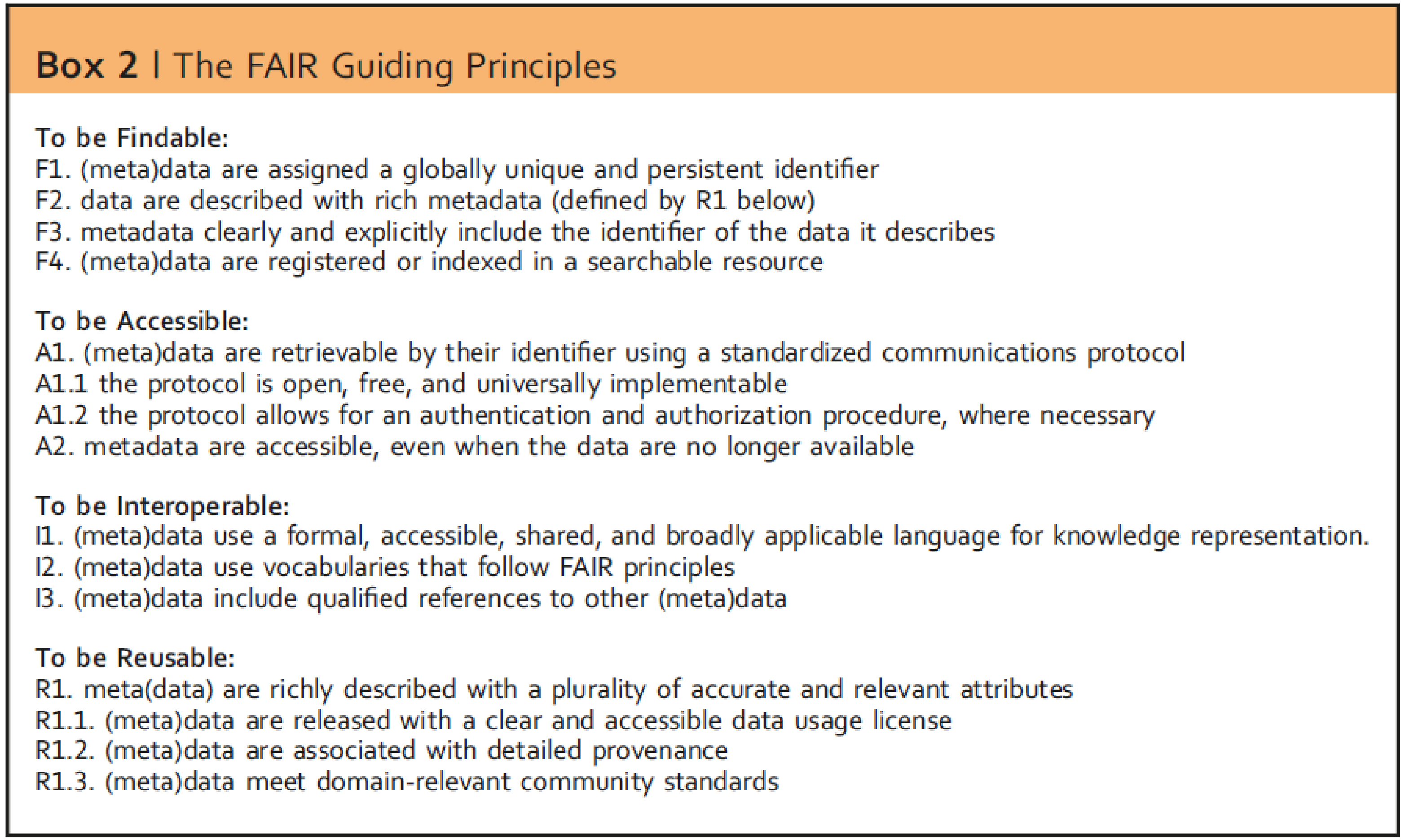
The FAIR principles

Key Thoughts:

4) The FAIR principles are not rules or standards.

The FAIR principles must not be mistaken for rules or standards that you can use to evaluate tools, data, policies, etc. This would soon make the principles out-of-date and inapplicable across research disciplines. Adopting the FAIR principles will often be a gradual adaptation of work routines – but it could also be a huge leap, where you replace one type of infrastructure with another. It will be up to the different research areas and research communities to make the FAIR principles work in their respective contexts.

The FAIR Guiding Principles



Box 2 | The FAIR Guiding Principles

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

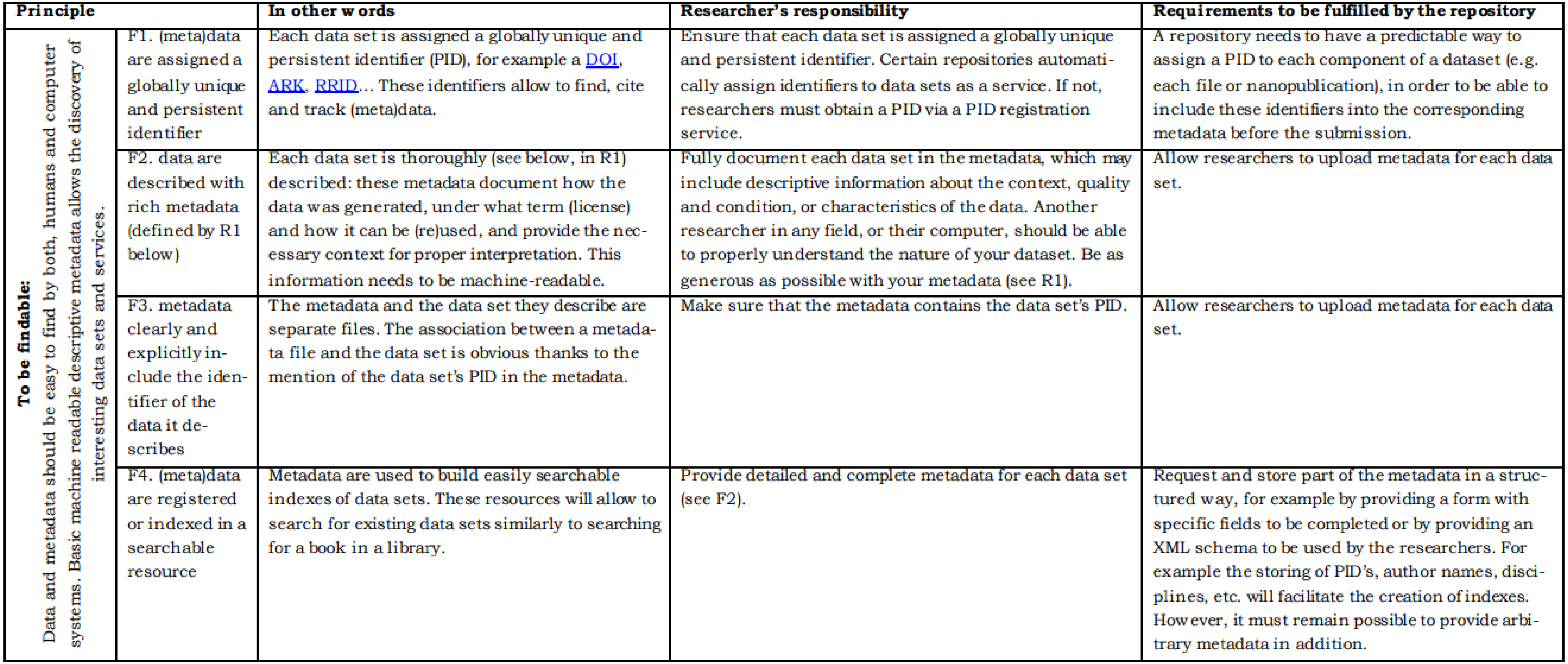
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even 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

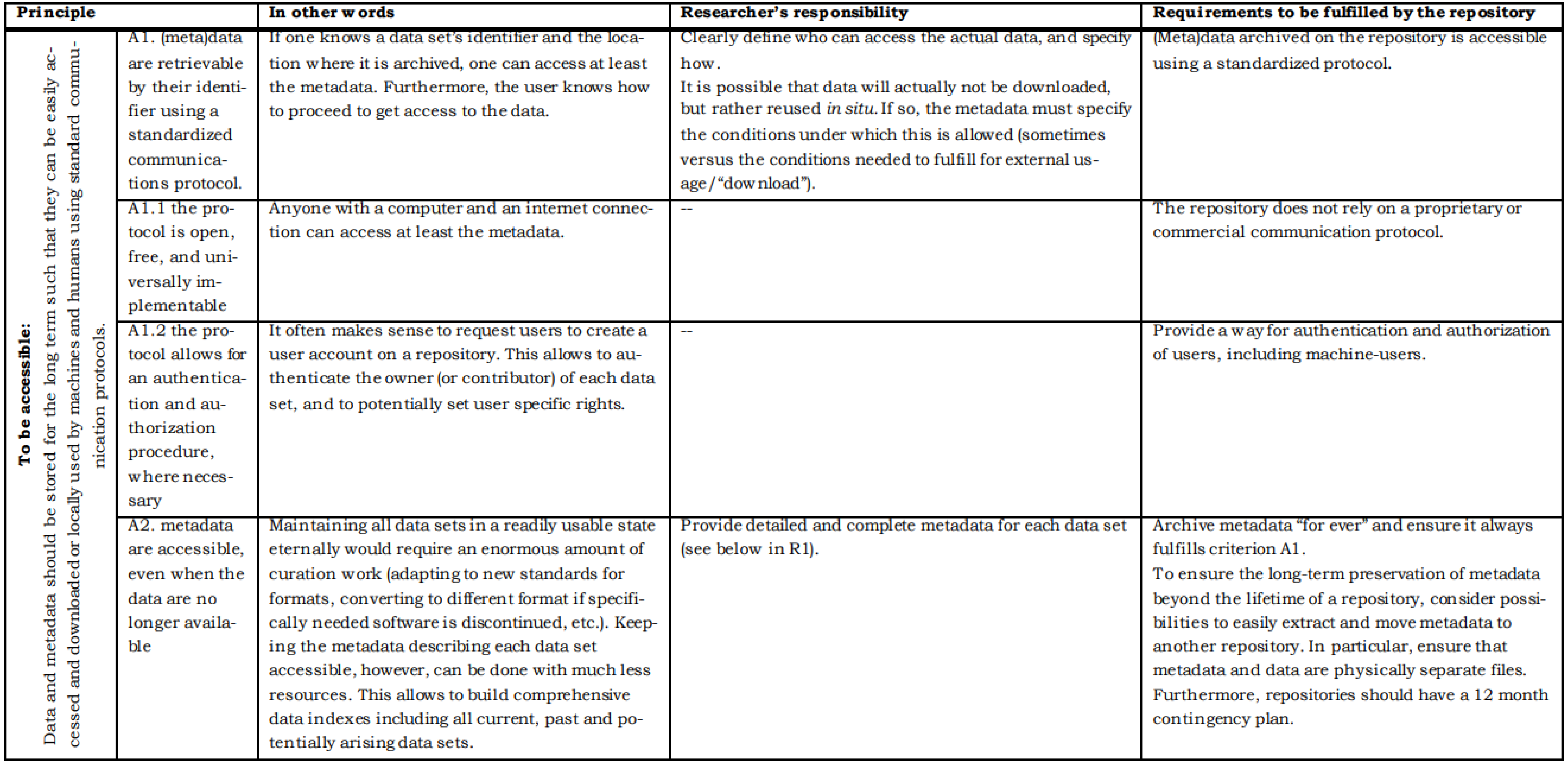
Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Swiss National Fund (SNF) and FAIR



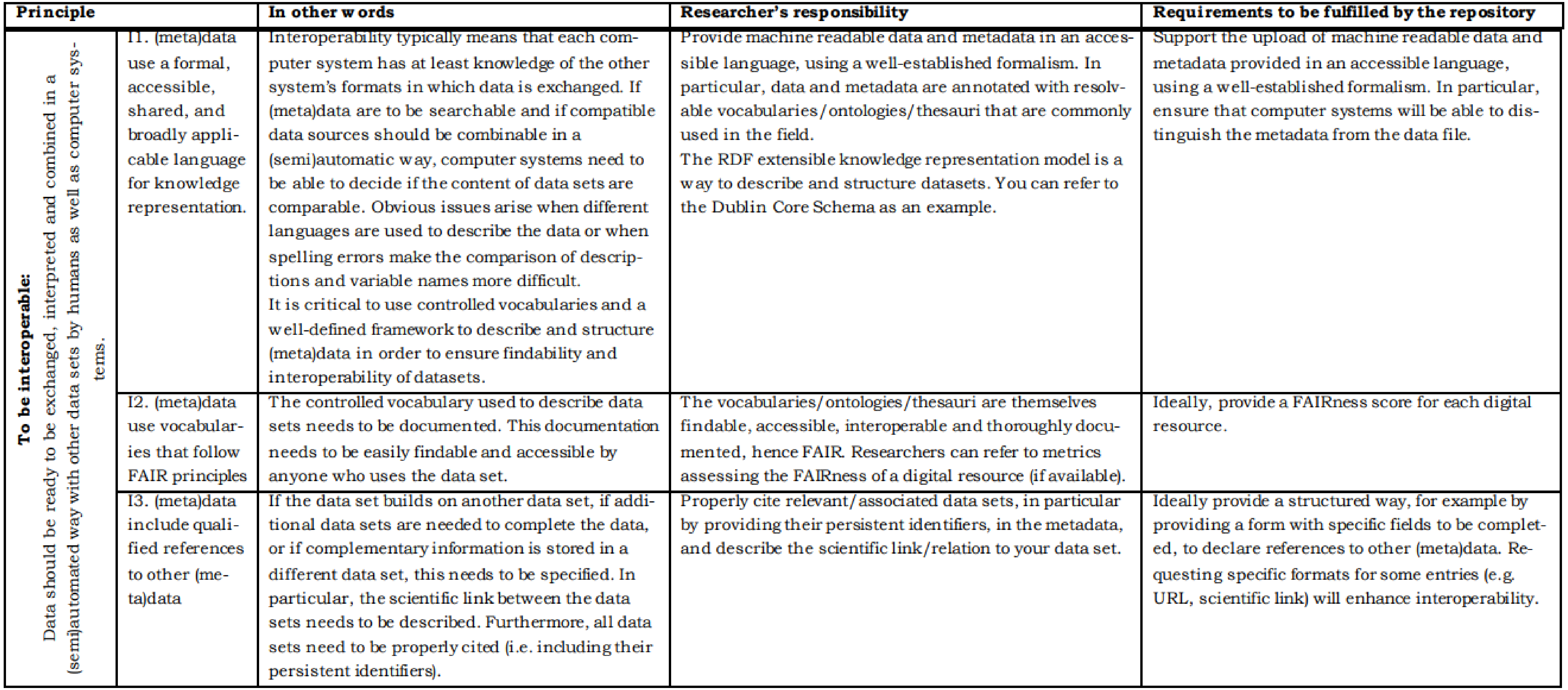
Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be findable: Data and metadata should be easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.	F1. (meta)data are assigned a globally unique and persistent identifier	Each data set is assigned a globally unique and persistent identifier (PID), for example a DOI, ARK, RRID… These identifiers allow to find, cite and track (meta)data.	Ensure that each data set is assigned a globally unique and persistent identifier. Certain repositories automatically assign identifiers to data sets as a service. If not, researchers must obtain a PID via a PID registration service.	A repository needs to have a predictable way to assign a PID to each component of a dataset (e.g. each file or nanopublication), in order to be able to include these identifiers into the corresponding metadata before the submission.
	F2. data are described with rich metadata (defined by R1 below)	Each data set is thoroughly (see below, in R1) described: these metadata document how the data was generated, under what term (license) and how it can be (re)used, and provide the necessary context for proper interpretation. This information needs to be machine-readable.	Fully document each data set in the metadata, which may include descriptive information about the context, quality and condition, or characteristics of the data. Another researcher in any field, or their computer, should be able to properly understand the nature of your dataset. Be as generous as possible with your metadata (see R1).	Allow researchers to upload metadata for each data set.
	F3. metadata clearly and explicitly include the identifier of the data it describes	The metadata and the data set they describe are separate files. The association between a metadata file and the data set is obvious thanks to the mention of the data set's PID in the metadata.	Make sure that the metadata contains the data set's PID.	Allow researchers to upload metadata for each data set.
	F4. (meta)data are registered or indexed in a searchable resource	Metadata are used to build easily searchable indexes of data sets. These resources will allow to search for existing data sets similarly to searching for a book in a library.	Provide detailed and complete metadata for each data set (see F2).	Request and store part of the metadata in a structured way, for example by providing a form with specific fields to be completed or by providing an XML schema to be used by the researchers. For example the storing of PID's, author names, disciplines, etc. will facilitate the creation of indexes. However, it must remain possible to provide arbitrary metadata in addition.

Swiss National Fund (SNF) and FAIR



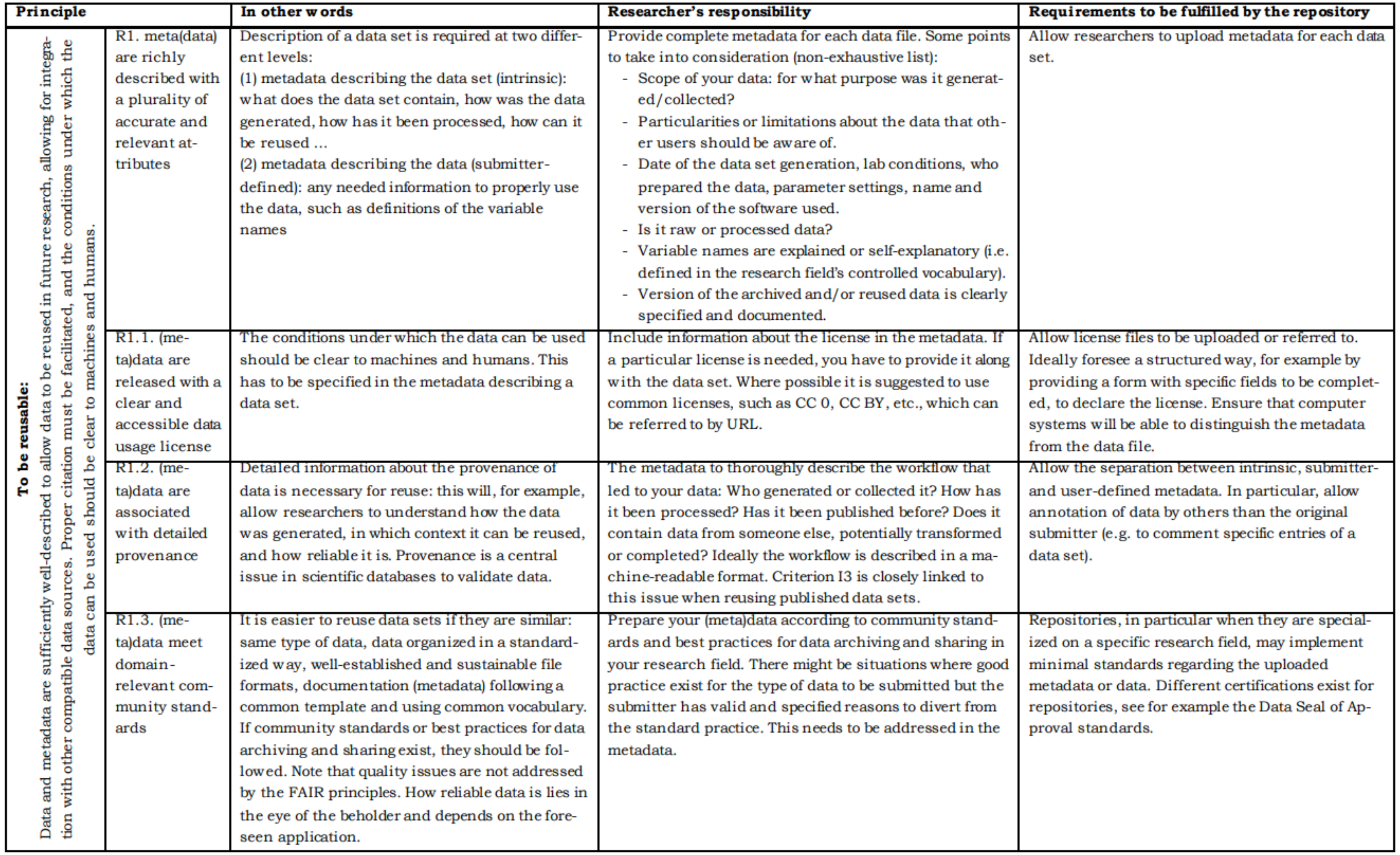
Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be accessible: Data and metadata should be stored for the long term such that they can be easily accessed and downloaded or locally used by machines and humans using standard communication protocols.	A1. (meta)data are retrievable by their identifier using a standardized communications protocol.	If one knows a data set's identifier and the location where it is archived, one can access at least the metadata. Furthermore, the user knows how to proceed to get access to the data.	Clearly define who can access the actual data, and specify how. It is possible that data will actually not be downloaded, but rather reused *in situ*. If so, the metadata must specify the conditions under which this is allowed (sometimes versus the conditions needed to fulfill for external usage/"download").	(Meta)data archived on the repository is accessible using a standardized protocol.
	A1.1 the protocol is open, free, and universally implementable	Anyone with a computer and an internet connection can access at least the metadata.	--	The repository does not rely on a proprietary or commercial communication protocol.
	A1.2 the protocol allows for an authentication and authorization procedure, where necessary	It often makes sense to request users to create a user account on a repository. This allows to authenticate the owner (or contributor) of each data set, and to potentially set user specific rights.	--	Provide a way for authentication and authorization of users, including machine-users.
	A2. metadata are accessible, even when the data are no longer available	Maintaining all data sets in a readily usable state eternally would require an enormous amount of curation work (adapting to new standards for formats, converting to different format if specifically needed software is discontinued, etc.). Keeping the metadata describing each data set accessible, however, can be done with much less resources. This allows to build comprehensive data indexes including all current, past and potentially arising data sets.	Provide detailed and complete metadata for each data set (see below in R1).	Archive metadata "for ever" and ensure it always fulfills criterion A1. To ensure the long-term preservation of metadata beyond the lifetime of a repository, consider possibilities to easily extract and move metadata to another repository. In particular, ensure that metadata and data are physically separate files. Furthermore, repositories should have a 12 month contingency plan.

Swiss National Fund (SNF) and FAIR



Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be interoperable: Data should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Interoperability typically means that each computer system has at least knowledge of the other system's formats in which data is exchanged. If (meta)data are to be searchable and if compatible data sources should be combinable in a (semi)automatic way, computer systems need to be able to decide if the content of data sets are comparable. Obvious issues arise when different languages are used to describe the data or when spelling errors make the comparison of descriptions and variable names more difficult. It is critical to use controlled vocabularies and a well-defined framework to describe and structure (meta)data in order to ensure findability and interoperability of datasets.	Provide machine readable data and metadata in an accessible language, using a well-established formalism. In particular, data and metadata are annotated with resolvable vocabularies/ontologies/thesauri that are commonly used in the field. The RDF extensible knowledge representation model is a way to describe and structure datasets. You can refer to the Dublin Core Schema as an example.	Support the upload of machine readable data and metadata provided in an accessible language, using a well-established formalism. In particular, ensure that computer systems will be able to distinguish the metadata from the data file.
	I2. (meta)data use vocabularies that follow FAIR principles	The controlled vocabulary used to describe data sets needs to be documented. This documentation needs to be easily findable and accessible by anyone who uses the data set.	The vocabularies/ontologies/thesauri are themselves findable, accessible, interoperable and thoroughly documented, hence FAIR. Researchers can refer to metrics assessing the FAIRness of a digital resource (if available).	Ideally, provide a FAIRness score for each digital resource.
	I3. (meta)data include qualified references to other (meta)data	If the data set builds on another data set, if additional data sets are needed to complete the data, or if complementary information is stored in a different data set, this needs to be specified. In particular, the scientific link between the data sets needs to be described. Furthermore, all data sets need to be properly cited (i.e. including their persistent identifiers).	Properly cite relevant/associated data sets, in particular by providing their persistent identifiers, in the metadata, and describe the scientific link/relation to your data set.	Ideally provide a structured way, for example by providing a form with specific fields to be completed, to declare references to other (meta)data. Requesting specific formats for some entries (e.g. URL, scientific link) will enhance interoperability.

Swiss National Fund (SNF) and FAIR



Principle		In other words	Researcher's responsibility	Requirements to be fulfilled by the repository
To be reusable: Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.	R1. meta(data) are richly described with a plurality of accurate and relevant attributes	Description of a data set is required at two different levels: (1) metadata describing the data set (intrinsic): what does the data set contain, how was the data generated, how has it been processed, how can it be reused … (2) metadata describing the data (submitter-defined): any needed information to properly use the data, such as definitions of the variable names	Provide complete metadata for each data file. Some points to take into consideration (non-exhaustive list): - Scope of your data: for what purpose was it generated/collected? - Particularities or limitations about the data that other users should be aware of. - Date of the data set generation, lab conditions, who prepared the data, parameter settings, name and version of the software used. - Is it raw or processed data? - Variable names are explained or self-explanatory (i.e. defined in the research field's controlled vocabulary). - Version of the archived and/or reused data is clearly specified and documented.	Allow researchers to upload metadata for each data set.
	R1.1. (meta)data are released with a clear and accessible data usage license	The conditions under which the data can be used should be clear to machines and humans. This has to be specified in the metadata describing a data set.	Include information about the license in the metadata. If a particular license is needed, you have to provide it along with the data set. Where possible it is suggested to use common licenses, such as CC 0, CC BY, etc., which can be referred to by URL.	Allow license files to be uploaded or referred to. Ideally foresee a structured way, for example by providing a form with specific fields to be completed, to declare the license. Ensure that computer systems will be able to distinguish the metadata from the data file.
	R1.2. (meta)data are associated with detailed provenance	Detailed information about the provenance of data is necessary for reuse: this will, for example, allow researchers to understand how the data was generated, in which context it can be reused, and how reliable it is. Provenance is a central issue in scientific databases to validate data.	The metadata to thoroughly describe the workflow that led to your data: Who generated or collected it? How has it been processed? Has it been published before? Does it contain data from someone else, potentially transformed or completed? Ideally the workflow is described in a machine-readable format. Criterion I3 is closely linked to this issue when reusing published data sets.	Allow the separation between intrinsic, submitter- and user-defined metadata. In particular, allow annotation of data by others than the original submitter (e.g. to comment specific entries of a data set).
	R1.3. (meta)data meet domain-relevant community standards	It is easier to reuse data sets if they are similar: same type of data, data organized in a standardized way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary. If community standards or best practices for data archiving and sharing exist, they should be followed. Note that quality issues are not addressed by the FAIR principles. How reliable data is lies in the eye of the beholder and depends on the foreseen application.	Prepare your (meta)data according to community standards and best practices for data archiving and sharing in your research field. There might be situations where good practice exist for the type of data to be submitted but the submitter has valid and specified reasons to divert from the standard practice. This needs to be addressed in the metadata.	Repositories, in particular when they are specialized on a specific research field, may implement minimal standards regarding the uploaded metadata or data. Different certifications exist for repositories, see for example the Data Seal of Approval standards.

So what exactly is FAIR data?

Findable

means that the data can be discovered by both humans and machines, for instance by exposing meaningful machine-actionable metadata and keywords to search engines and research data catalogues. The data are referenced with unique and persistent identifiers (e.g. DOIs or Handles) and the metadata include the identifier of the data they describe.

So what exactly is FAIR data?

Accessible

means that the data are archived in long-term storage and can be made available using standard technical procedures. This does not mean that the data have to be openly available for everyone, but information on how the data could be retrieved (or not) has to be available. For example, data can be marked “Access only with explicit permission from the author” and include the author’s contact details. Ideally, though, the information about data accessibility can also be read by machines, e.g. by way of machine-readable standard licences.

So what exactly is FAIR data?

Interoperable

means that the data can be exchanged and used across different applications and systems — also in the future, for example, by using open file formats. It also means that the data can be integrated with other data from the same research field or data from other research fields. This is made possible by using metadata standards, standard ontologies, and controlled vocabularies as well as meaningful links between the data and related digital research objects.

So what exactly is FAIR data?

Reusable

means that the data are well documented and curated and provide rich information about the context of data creation. The data should conform to community standards and include clear terms and conditions on how the data may be accessed and reused, preferably by applying machine-readable standard licenses. This allows others either to assess and validate the results of the original study, thus ensuring *data reproducibility*, or to design new projects based on the original results, in other words *data reuse* in the stricter sense. Reusable data encourage collaboration and avoid duplication of effort.

Why FAIR

Why use the FAIR principles for your research data?

Reusing existing data sets for new research purposes is becoming more common across all research disciplines.

Research funders and publishers are asking researchers to make data sets produced in their projects available to others. And research institutes are promoting measures to secure the transparency and accessibility of locally produced data sets. To facilitate this, datasets need to be Findable, Accessible, Interoperable and Reusable.

This is what the FAIR principles are all about.

Why FAIR

Why FAIR?

□ Help peers and your future-self understand the research project and data

□ Facilitate data sharing and collaborations

□ Increase the visibility of research and can lead to more citations

□ Improve the transparency, reliability and reproducibility of research

□ Prevent data loss

And thereby:

□ Maximize potential from data assets

□ Maximize research impact

FAIRification practices

How you apply the FAIR principles, depends on your specific discipline and your way of doing research. But there are different activities you must consider within your research workflows, if you want to make your data FAIR.

For instance:

✓ documenting your data

✓ choosing appropriate file formats

✓ adding metadata

✓ giving access to the data

✓ licensing the data or adding a persistent identifier

How FAIR?

FAIRification practices

Findable

- Metadata
- PIDs
- Repositories

Accessible

- Metadata
- Open file formats and software

Interoperable

- Metadata
- Ontologies
- Repositories

Reusable

- Metadata
- Licences

How FAIR?

FAIRification practices

FAIR ≠ Open

as open as possible, as closed as necessary

Image: 'Balancing rocks' by ViewminderCC-BY-SA-ND www.flickr.com/photos/light_seeker/7780857224

D.B. Deutz, M.C.H. Buss, J. S. Hansen, K. K. Hansen, K.G. Kjelmann, A.V. Larsen, E. Vlachos, K.F. Holmstrand (2020). How to FAIR: a Danish website to guide researchers on making research data more FAIR
https://doi.org/10.5281/zenodo.3712065

How FAIR?

How to Fair

Documentation

Documentation adds context to your data and makes the data easier to understand and reuse in the future.

File Formats

File formats determine how data can be used. It is important to decide what file formats to use for data collection, data processing, data archiving, and long-term preservation.

Access to Data

Access to data means that you determine who you make your data available for, how you provide access, and under which conditions.

Persistent Identifiers

To make your data easy to find and accessible, you must provide your data and metadata with a persistent identifier. A persistent identifier is a long-lasting reference to a digital resource and provides the information required to reliably identify, verify and locate your research data

Data Licenses

A data license is a legal arrangement between the creator of the data and the end-user specifying what users can do with the data.

FAIR - Documentation

What is documentation?

Imagine finding a dataset you created a long time ago. Now think of the contextual information that you would need to determine whether these data are relevant to your current research and whether you would be able to understand how they were created:

- What data?
- What data type?
- Who created the data?
- When?
- Where?
- In which context?
- By which method?
- … and so on.

FAIR - Documentation

Two general types of data documentation

Data-level documentation

Data-level documentation includes information about specific data files like:
- Data type
- Structure of the data, e.g. questions, variables, concepts
- Data processing procedures, … and so on (this list is not exhaustive)

Project-level documentation

Project- or study-level documentation describes:
- When, how and why the data were generated and by whom
- How the data were processed
- What quality assurance measures have been used, … and so on (this list is not exhaustive)

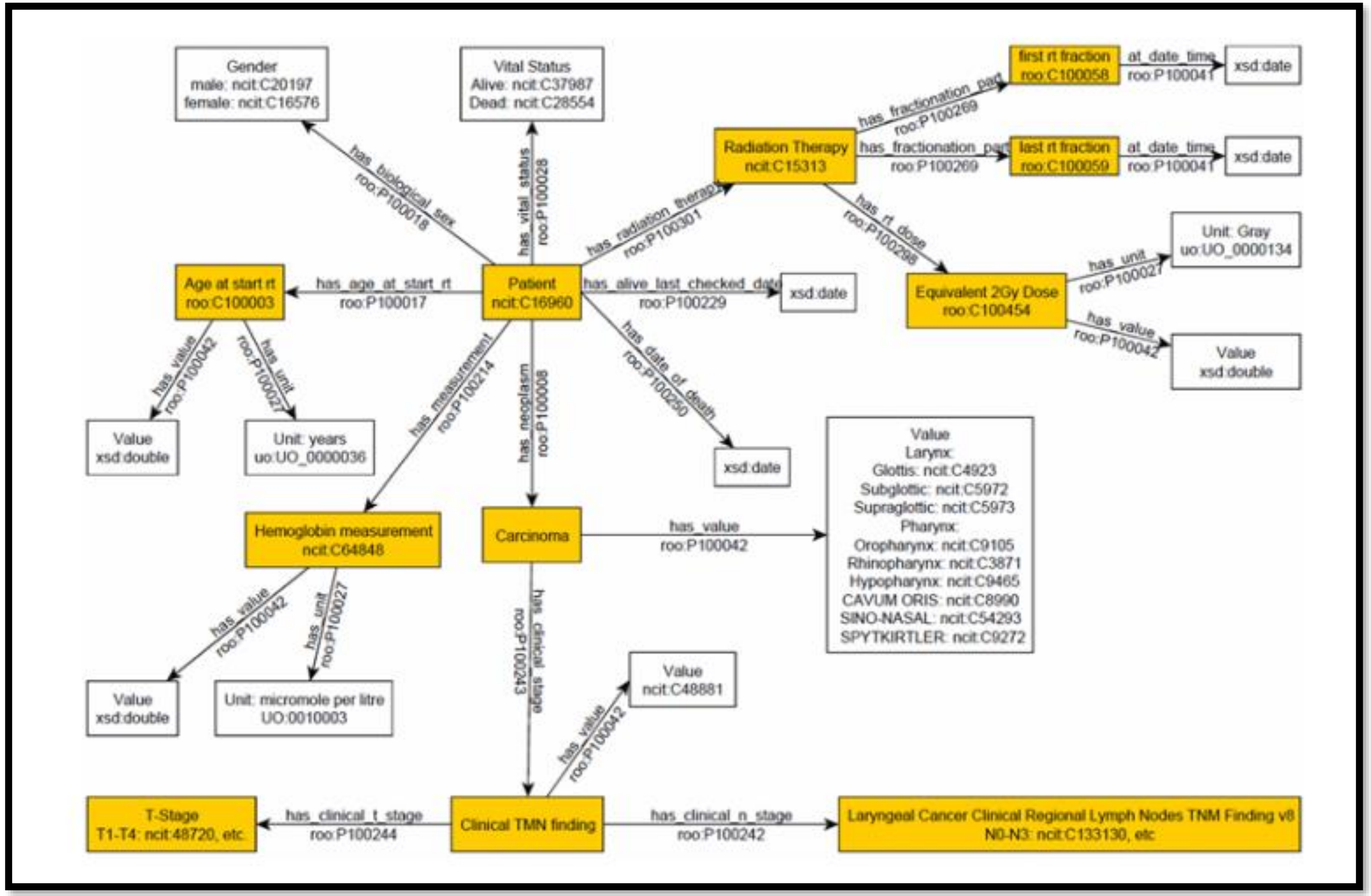
How FAIR?

FAIR - Documentation

Example: **Data map**

For small projects the entire code is stored.

For larger projects the researchers prefer to describe the method, the model selection and the packages used.



FAIR - Documentation

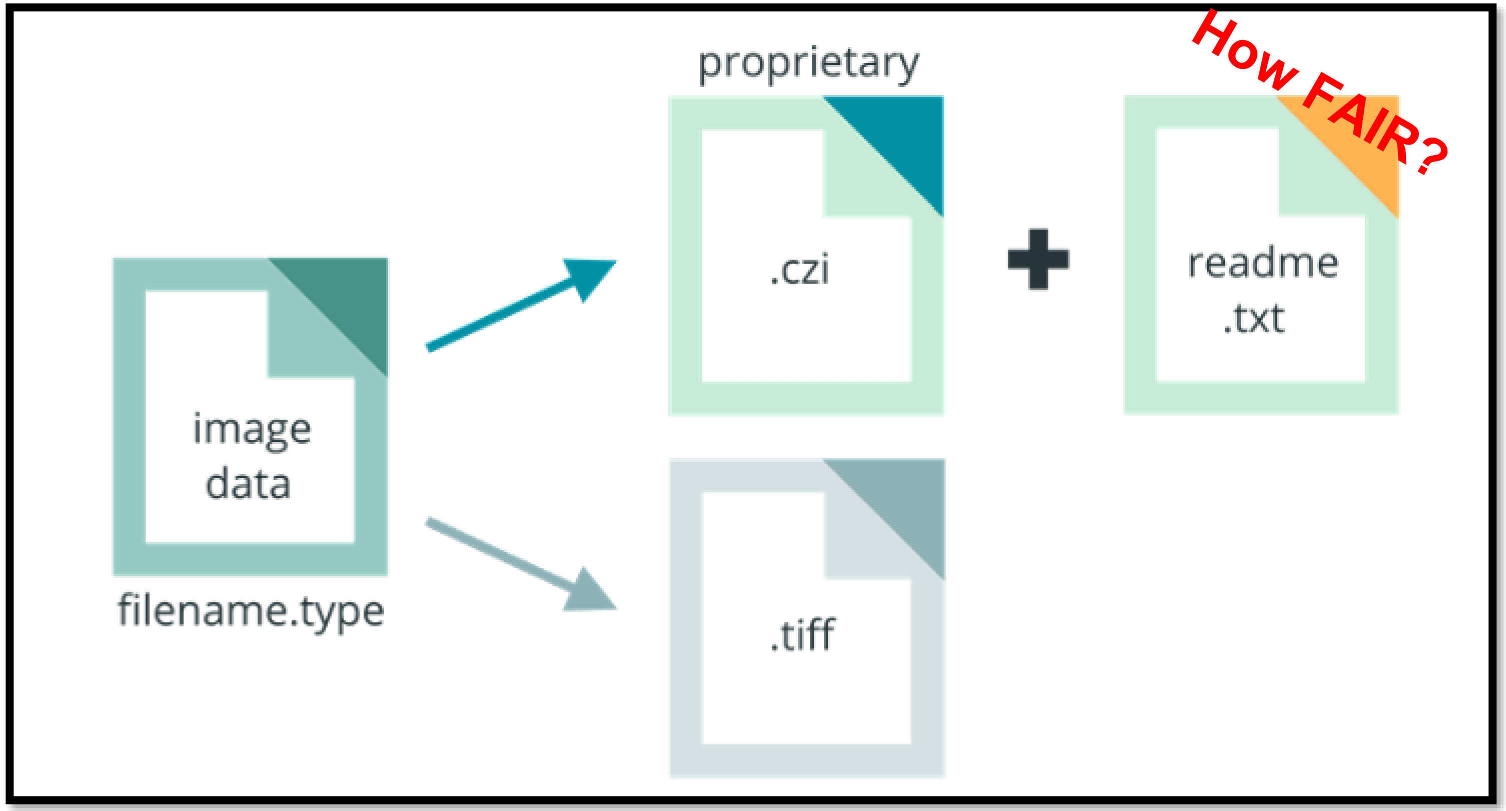
Publish and preserve

FAIR documentation is what enables you as a researcher to show how the data was generated and for what purpose. Think about what information is necessary for this to happen:

- ❑ Methodology descriptions
- ❑ Codebooks
- ❑ Questionnaires
- ❑ Scripts like editor- and do-files (STATA)
- ❑ Laboratory notebooks and experimental protocols
- ❑ Software syntax and output files
- ❑ Database schemes
- ❑ Provenance information about secondary data

FAIR - File Formats

Files are usually named like this: **[prefix].[suffix]** or filename.type. In this way files of the type .txt are text encoded files and usually contain text and/or numbers. Images are often saved in .jpg or .bmp while audio can be saved in the .mp3 or .wav file format.



Some file formats are **proprietary** – like .nef or .wma which are owned by Nikon and Microsoft. Other file formats like .txt or .csv are **non-proprietary** and can be used with a variety of software. The purpose of a file should help determine which file format to choose. Therefore, you may have to keep some data files in multiple formats. It is important to plan what file formats to use for each purpose: data collection/ processing/analysis, reuse, and preservation.

FAIR - File Formats

Some examples of preferred FAIR file formats for preservation

Containers:	TAR, GZIP, ZIP
Databases:	XML, CSV, JSON
Geospatial:	SHP, DBF, GeoTIFF, NetCDF
Video:	MPEG, AVI, MXF, MKV
Sounds:	WAVE, AIFF, MP3, MXF, FLAC
Statistics:	DTA, POR, SAS, SAV
Images:	TIFF, JPEG 2000, PDF, PNG, GIF, BMP, SVG
Tabular data:	CSV, TXT
Text:	XML, PDF/A, HTML, JSON, TXT, RTF
Web archive:	WARC

FAIR - Metadata

From a FAIR perspective, metadata are more important than your data, because metadata would always be openly available and they link research data and publications in the ***Internet of FAIR Data and Services***. The distinction between data and metadata is not ontological, but it is grounded in use. What is “data” and what is “metadata” is thereby a matter of perspective: Some researchers’ metadata can be other researchers’ data.

While data documentation is meant to be read and understood by humans, metadata (which are sometimes a part of the documentation) are primarily meant to be processed by machines.

FAIR - Metadata

Administrative metadata are data about a project or resource that are relevant for managing it; for example, project/ resource owner, principal investigator, project collaborators, funder, project period, etc. They are usually assigned to the data, before you collect or create them.

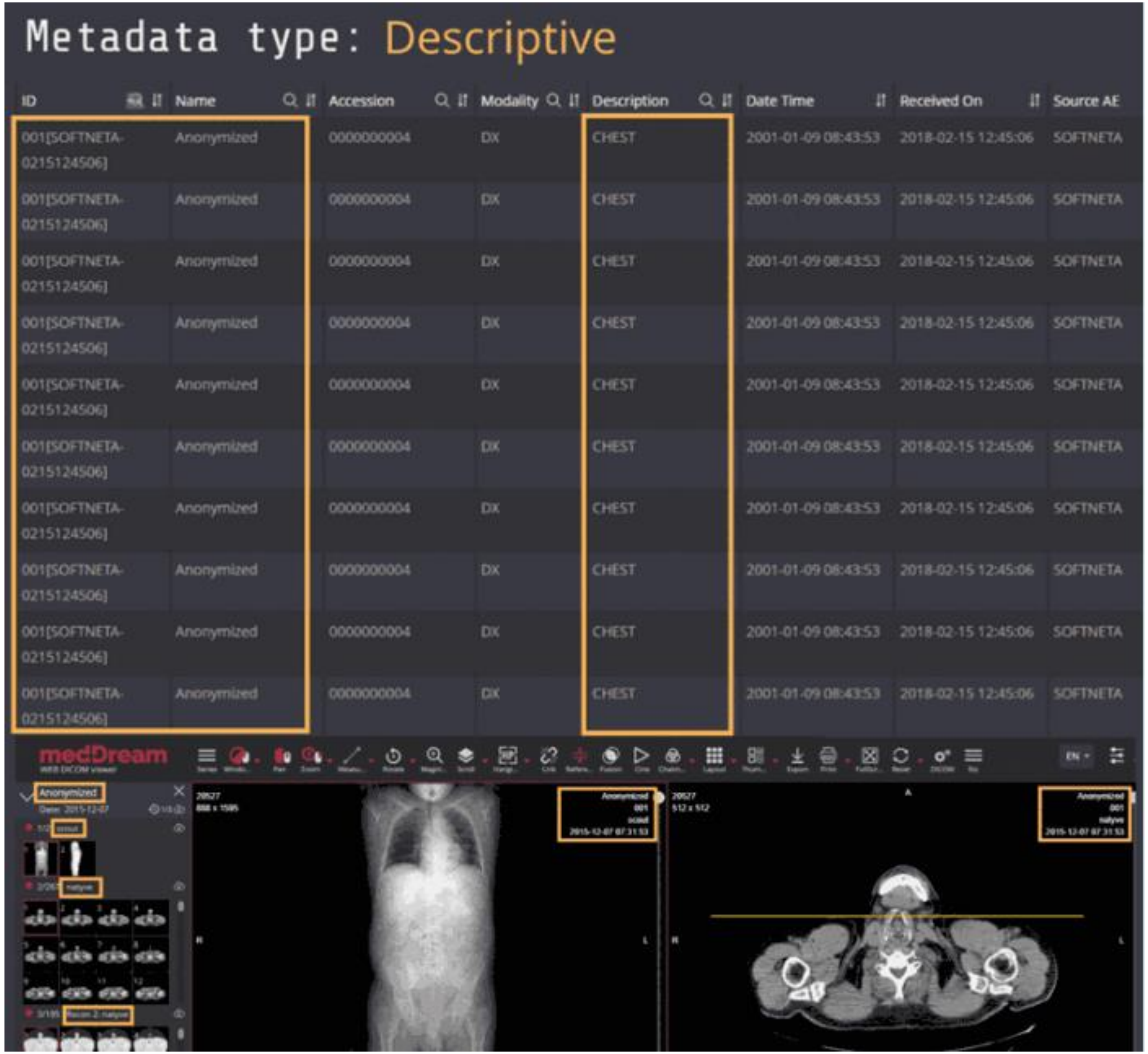
Descriptive or citation metadata are data about a dataset or resource that allow people to discover and identify it; for example, authors, title, abstract, keywords, persistent identifier, related publications, etc.

Structural metadata are data about how a dataset or resource came about, but also how it is internally structured. Structural metadata describe, for example, the unit of analysis, collection method, sampling procedure, sample size, categories, variables, etc. Structural metadata have to be gathered by the researchers according to best practice in their research community and will be published together with the data. Descriptive and structural metadata should be added continuously throughout the project.

How FAIR?

FAIR - Metadata

Descriptive or citation metadata



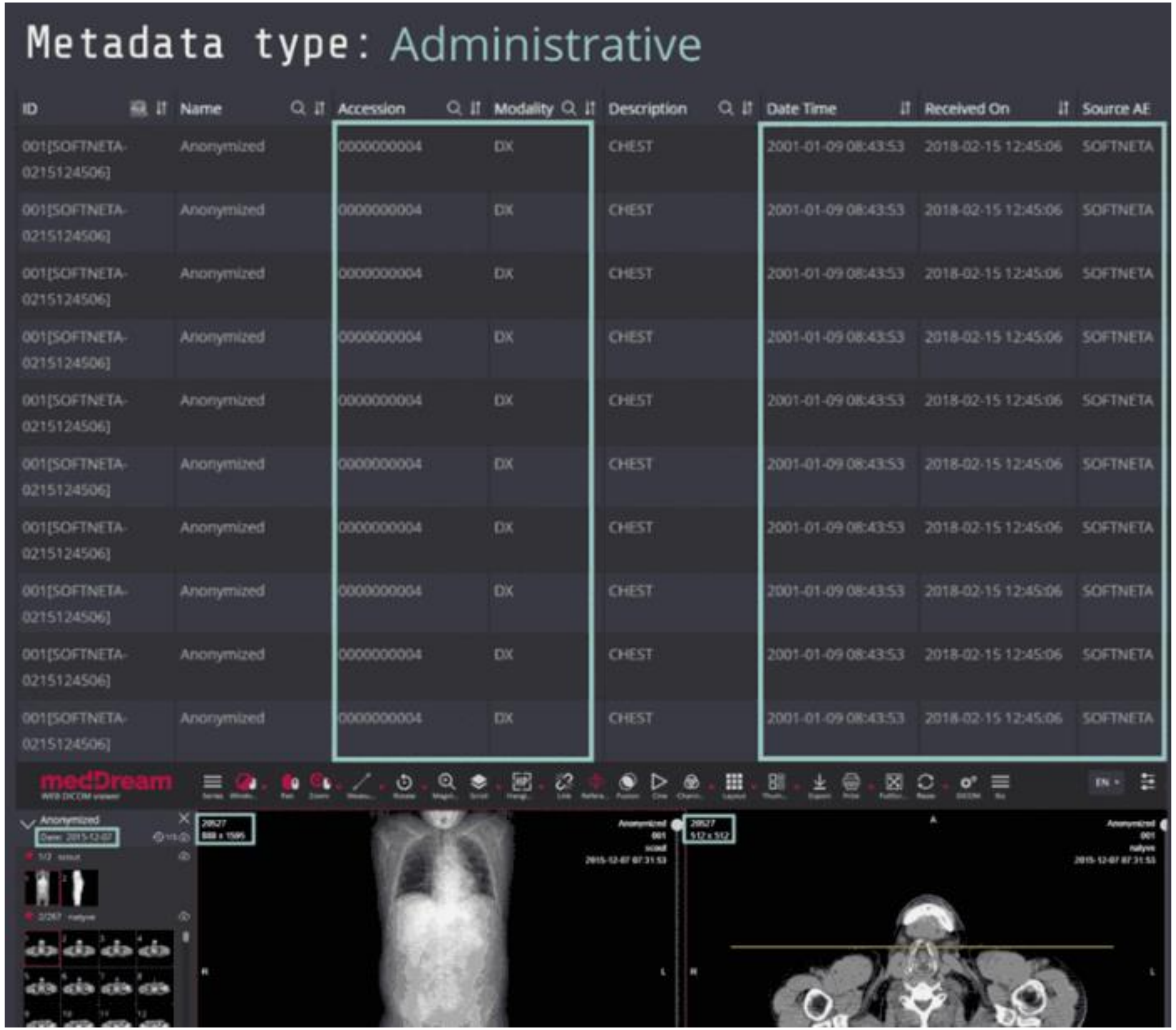
Metadata type: Descriptive

ID	Name	Accession	Modality	Description	Date Time	Received On	Source AE
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA

How FAIR?

FAIR - Metadata

Administrative metadata

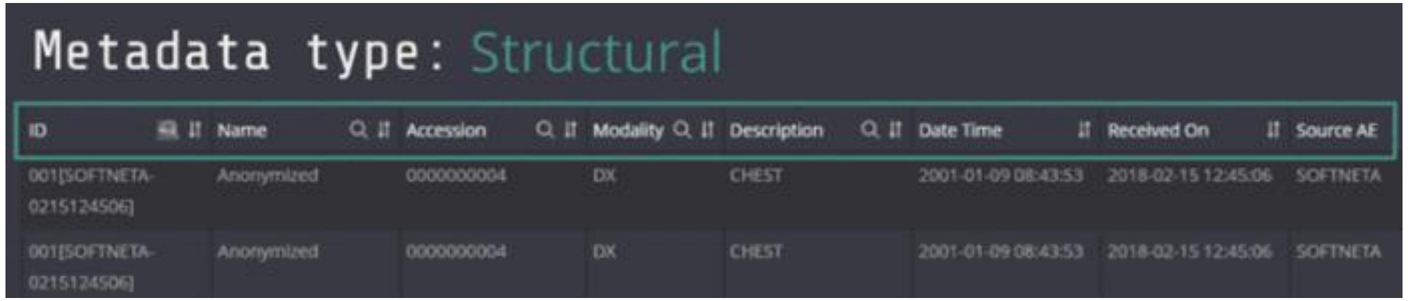


Metadata type: Administrative

ID	Name	Accession	Modality	Description	Date Time	Received On	Source AE
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA

FAIR - Metadata

Structural metadata



Metadata type: Structural

ID	Name	Accession	Modality	Description	Date Time	Received On	Source AE
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA
001[SOFTNETA-0215124506]	Anonymized	0000000004	DX	CHEST	2001-01-09 08:43:53	2018-02-15 12:45:06	SOFTNETA

FAIR - Metadata

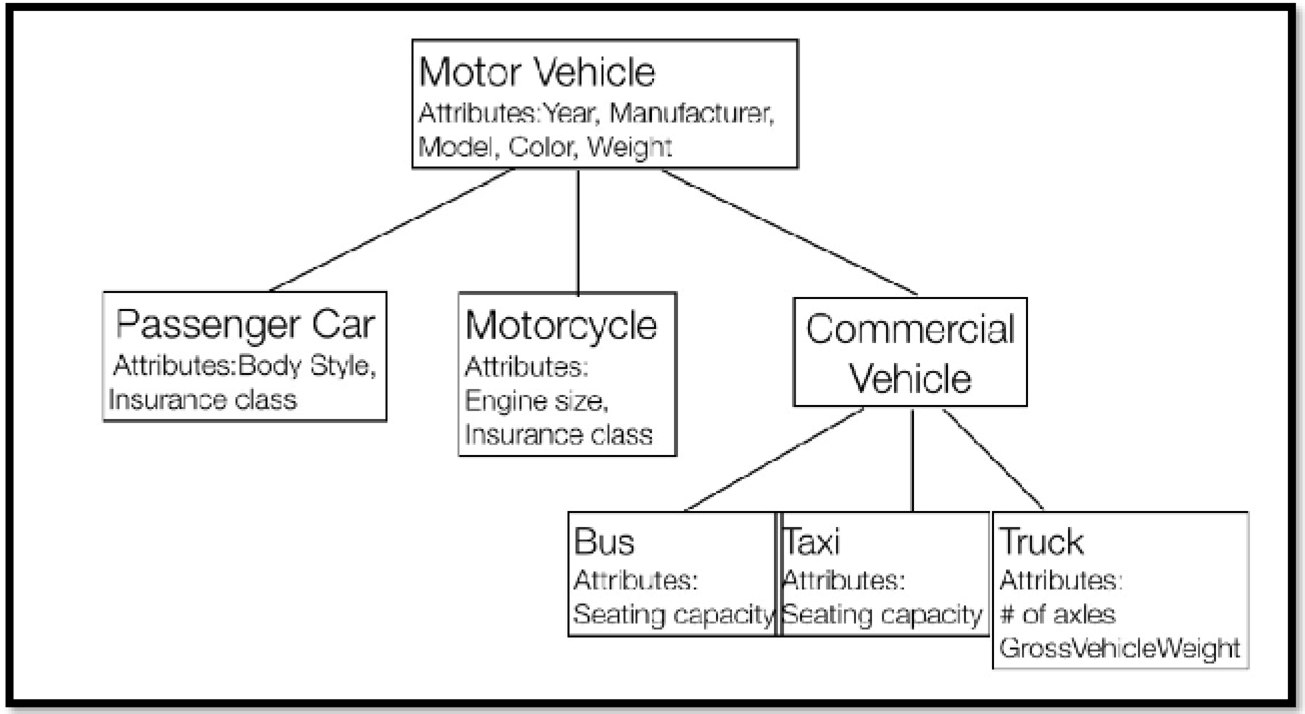
A metadata **standard** is a subject-specific guide to your metadata. Metadata elements are grouped into sets designed for a specific purpose and given a standard name and definition. Rules on what content must be included, what syntax must be used, or a controlled vocabulary can also be included in a metadata standard. A starting point can be a taxonomy, or an ontology.

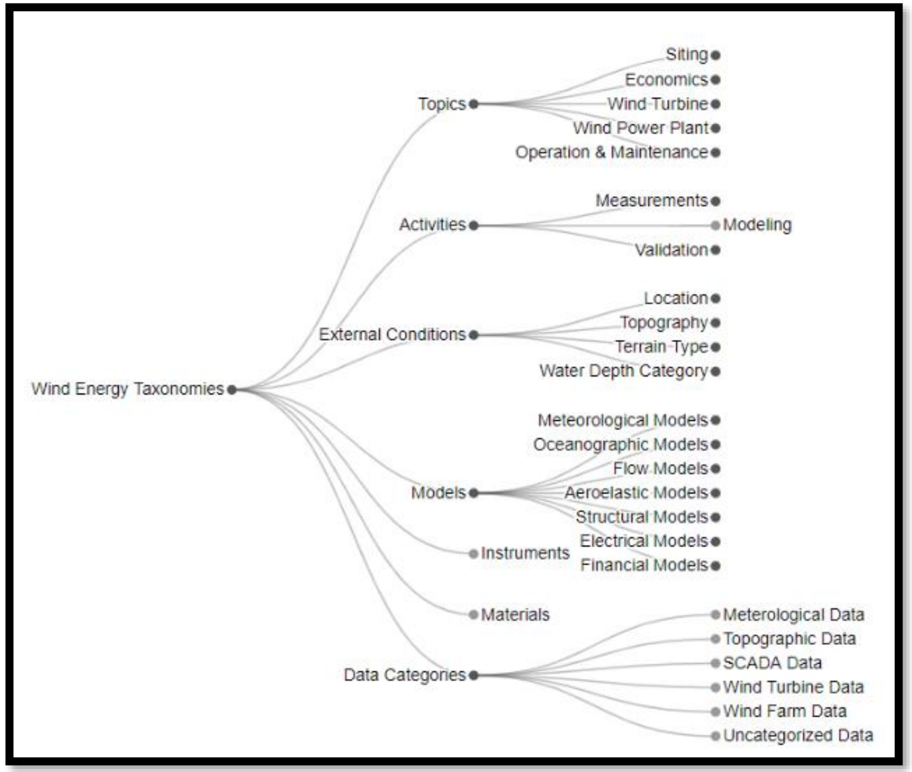
Does a taxonomy or ontologies exist in your field of work/research?

How FAIR?

FAIR - Metadata

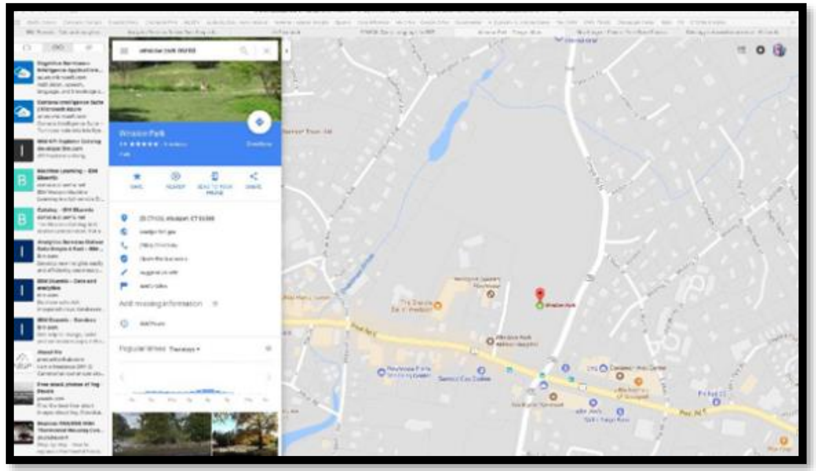
Taxonomy vs Ontology

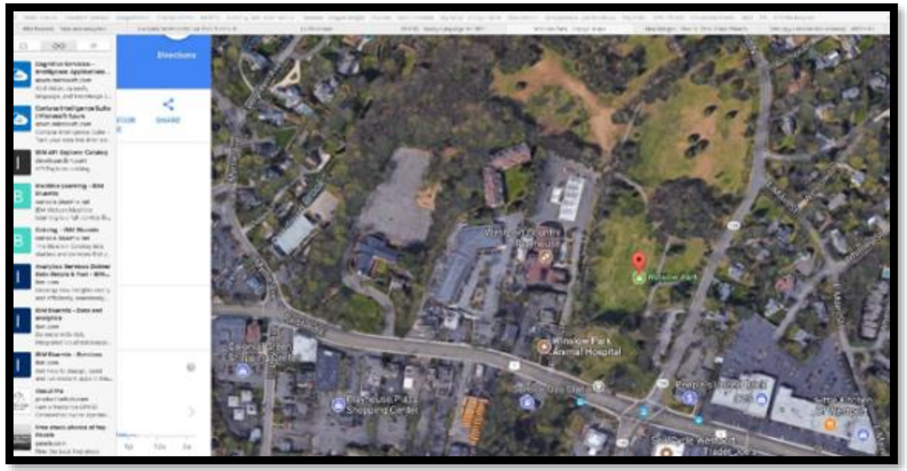

Motor Vehicle
Attributes:Year, Manufacturer, Model, Color, Weight
Passenger Car
Attributes:Body Style, Insurance class
Motorcycle
Attributes: Engine size, Insurance class
Commercial Vehicle
Bus
Attributes: Seating capacity
Taxi
Attributes: Seating capacity
Truck
Attributes: # of axles GrossVehicleWeight



FAIR - Metadata

Taxonomy vs **Ontology**





"All three maps or domains contain Winslow Park and in a global sense, could be in the same Taxonomy. But these different domains or ontologies have very specific uses. For example, a history teacher lecturing on the history of Winslow park in the United States, may find the first map more useful."

https://www.dataversity.net/taxonomy-vs-ontology-machine-learning-breakthroughs/

How FAIR?

FAIR – Access to Data

…. In short … To search for a suitable repository for your research data you can visit re3data.org, which is a global registry of research data repositories from different academic disciplines, FAIRsharing, which allows you to discover databases grouped by domain, species or organization, or check the links page to find more resources on data repositories.

FAIR – Persistent Identifiers

What is a persistent identifier?

A persistent identifier (PID) is a long-lasting reference to a digital resource and provides the information required to reliably identify, verify and locate your research data eliminating many misunderstandings. A PID may also be connected to a set of metadata which describes a digital resource.

Notable persistent identifiers are the Digital Object Identifier (**DOI**) and the Handle System which can both be assigned to data to identify them uniquely. The DOI system uses the Handle System, which is the best infrastructure component available today for managing digital objects. While DOIs are mainly assigned to resources ready for public dissemination, **Handles** are in general used to persistently identify other categories of digital resources (e.g. those created in the labs) to make them referable by software, workflows etc.

FAIR – Persistent Identifiers

How to get one for your data?

- ❑ Browsing through the list of repositories recommended by the **European Research Council**.
- ❑ Visiting **re3data.org**, which is a global registry of research data repositories from different academic disciplines.
- ❑ Exploring **FAIRsharing**, which allows you to discover databases grouped by domain, species or organization.
- ❑ Or check our recommended data repositories listed **here**.

FAIR – Persistent Identifiers

Information about Persistent Identifiers (PID)

DOI: List of current DOI registration agencies provided by the International DOI Foundation

Handle: Assigning, managing and resolving persistent identifiers for digital objects and other Internet resources provided by the Corporation for National Research Initiatives (CNRI)

PURL: Persistent Identifiers developed by the Online Computer Library Center (OCLC). Since 2016 hosted by the Internet Archive

URN: List of all registered namespaces provided by the Internet Assigned Numbers Authority (IANA)

Nice to Know

Nice to Know

FAIR - Copyright

What is copyright, who owns it and how long does it last?

Copyright is an intellectual property right assigned automatically to the creator. It prevents unauthorised copying and publishing of an original work. Copyright applies to research data and plays a role when creating, sharing and reusing data.

https://ukdataservice.ac.uk/learning-hub/research-data-management/rights-in-data/copyright/

Nice to Know

FAIR - Copyright

What is covered?

Under the Copyright, Designs and Patents Act 1988 copyright applies to:

Original literary, dramatic, musical or artistic works.

Sound recordings, films, broadcasts or cable programmes.

The typographical arrangement of publications.

Most research outputs, such as spreadsheets, publications, reports and computer programs, fall under literary work and are therefore protected by copyright. Facts, however, cannot be copyrighted.

Nice to Know

FAIR - Copyright

Useful facts about copyright ownership and transfer

- ❑ The author(s) or creator(s) of a work automatically own(s) copyright and this can be assumed as soon as the work exists in a recorded form.
- ❑ For copyright to apply, the work must be original and fixed in a material form (written or recorded); there is no copyright in ideas or unrecorded speech.
- ❑ If a work has two authors, the copyright will by default be owned by both authors.
- ❑ For work created during employment, legally, the copyright owner is the employer, subject to ‘any agreement to the contrary’. In practice, many academic institutions assign copyright in research materials and publications to the researchers, but researchers should check how their institution assigns copyright.

Nice to Know

FAIR - Copyright

Useful facts about copyright ownership and transfer

- ❑ For collaborative research or derived data, copyright is held by all the investigators or institutions involved.
- ❑ For data collected via interviews that are recorded and/or transcribed, the researcher holds the copyright of recordings and transcripts but each speaker is an author of his or her recorded words in the interview (Padfield, T (2010) *Copyright for archivists and records managers, 4th ed.*, London: Facet Publishing).
- ❑ Copyright can be transferred by the owner but only in writing by means of a transfer document called an assignment.
- ❑ If researchers wish to publish large extracts from an interview, it is advisable to obtain a transfer of copyright from interviewees.

Nice to Know

FAIR - Copyright

Useful facts about copyright ownership and transfer

❑ Creators of a work can also hold moral rights and publications rights.

❑ A database may be protected by copyright in the content and database right in the structure.

❑ Data can be reproduced for non-commercial teaching or research purposes without infringing copyright under the fair dealing concept, providing that the data source used, data distributor and the copyright holder are acknowledged

❑ If secondary users wish to reproduce data, they must obtain copyright clearance from the rights holder.

Nice to Know

FAIR - Copyright

Cite the Data

Citing a dataset correctly is just as important as citing articles, books, images and websites – each dataset is a source of evidence to support your argument.

Reasons to Cite the Data

- ✓ Support reproducibility of your research and attribute credit to the researcher.
- ✓ Support data creators – we assign data a Digital Object Identifier (DOI) free of charge.
- ✓ Make identifying and finding data easier. The DOI will always link to the data, even when its location changes.
- ✓ Citations enable tracking, measuring of impact, demonstrating use and value to funders and potential refunding.
- ✓ Funding bodies encourage the research community to establish data citation as the rule rather than the exception.

Nice to Know

FAIR - Copyright

Copyright vs **Copyleft**

What is "copyleft"? Is it the same as "open source"?

"Copyleft" refers to licenses that allow derivative works but require them to use the same license as the original work. For example, if you write some software and release it under the GNU General Public License (a widely-used copyleft license), and then someone else modifies that software and distributes their modified version, the modified version must be licensed under the GNU GPL too — including any new code written specifically to go into the modified version. Both the original and the new work are Open Source; the copyleft license simply ensures that property is perpetuated to all downstream derivatives.

https://opensource.org/faq#copyleft

Nice to Know

FAIR - Copyright

Other rights to be aware of beyond copyright

Moral rights

Besides copyright, the creator of a work also holds moral rights, which cannot be transferred. Moral rights give the creator the right to be identified as the author of a work. This right must be asserted by the author in writing.

Moral rights typically last the same length of time as copyright. Unlike copyright they cannot be sold to another party but can be waived by the author. They can also be bequeathed upon an author's death.

Nice to Know

FAIR - Copyright

Other rights to be aware of beyond copyright

Publication right

Publication right offers rights equivalent to copyright to a person who publishes previously unpublished material that has already fallen out of copyright.

This right rewards the creative effort employed in editing another research work. Creating and publishing a database based on unpublished historical source material, which is not in copyright, will give the creator a publication right.

Nice to Know

FAIR - Copyright

Other rights to be aware of beyond copyright

Database rights

If information is structured in a database, the structure acquires a database right, alongside the copyright in the content of the database. Legally, a database is a collection of independent works arranged in a systematic or methodical way.

A database may be protected by both copyright and database right. For a database right to apply, the database must be the result of substantial intellectual investment in obtaining, verifying or presenting the content in an original manner. Simply entering facts into a spreadsheet does not count as substantial effort. The database right is an automatic right and protects databases against the unauthorised extraction and reuse of the contents.

Database rights are protected for 15 years from the date of creation or publication. For some complex databases, the structure itself can be categorised as a literary work (even if its contents are of a visual nature) and attract 70 years' copyright similar to other literary material.

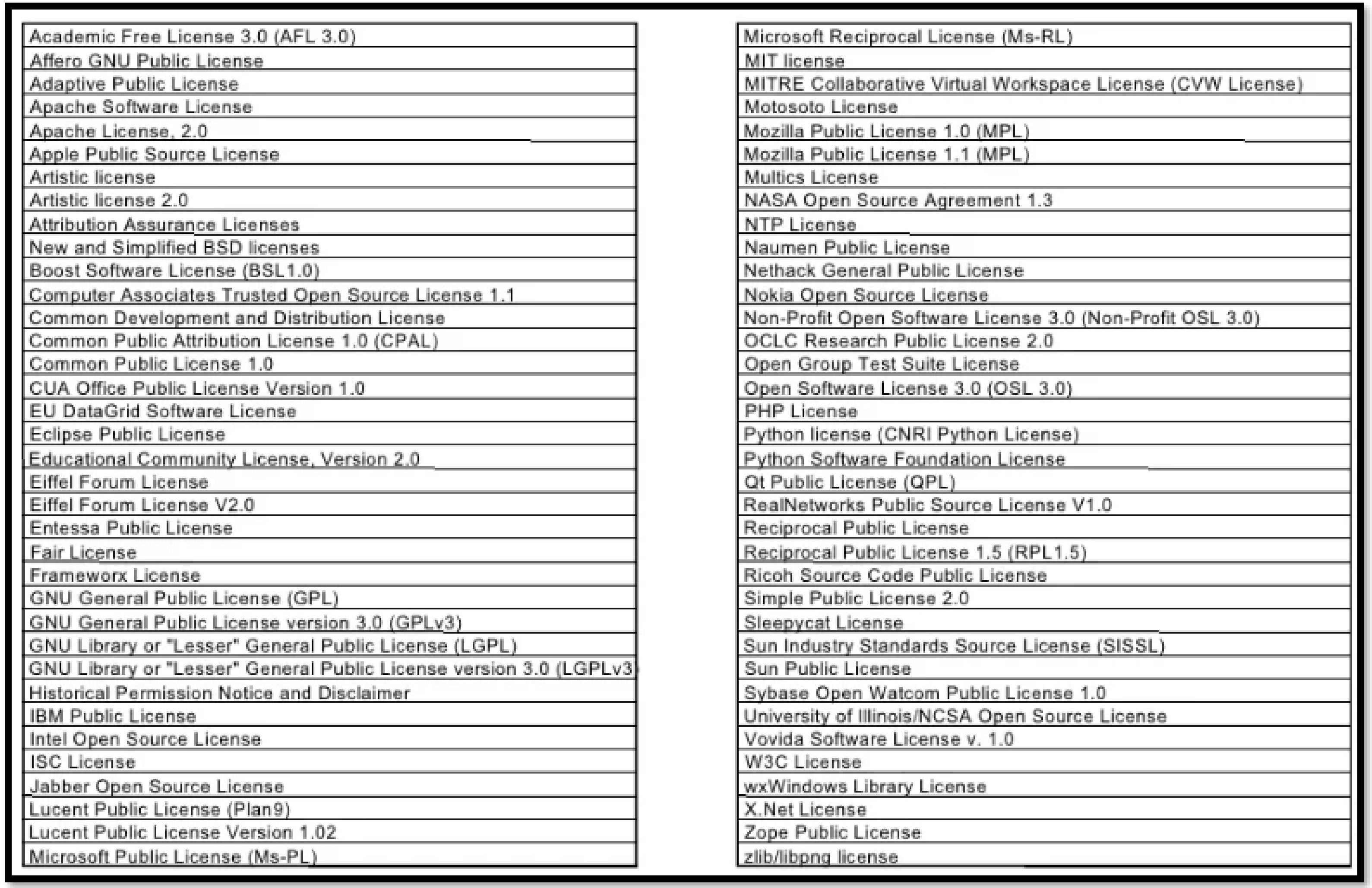
License Agreement

Nice to Know

FAIR - Licencing

Software vs. Data

Open Source Initiative
– approved:



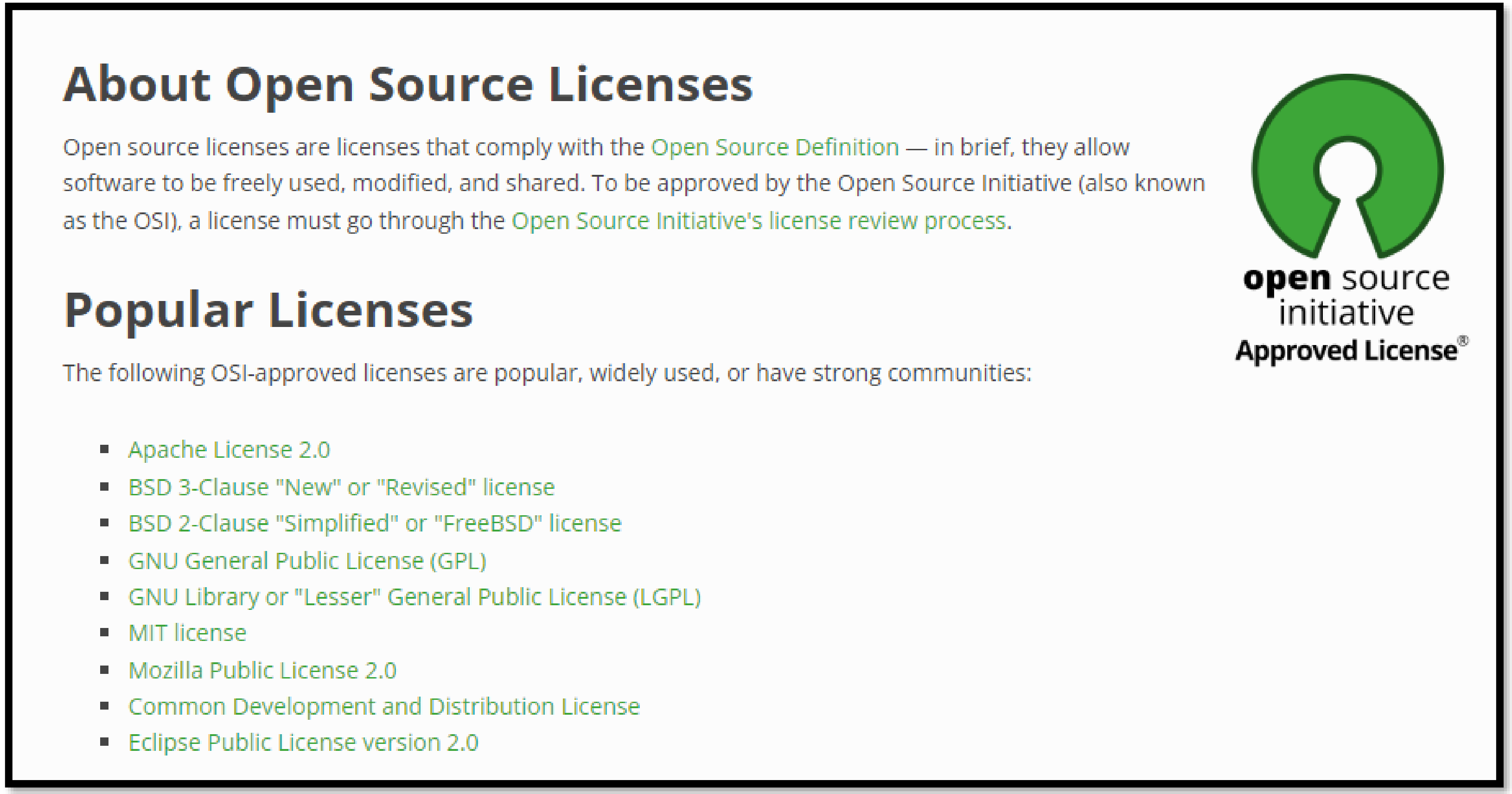
Academic Free License 3.0 (AFL 3.0)
Affero GNU Public License
Adaptive Public License
Apache Software License
Apache License, 2.0
Apple Public Source License
Artistic license
Artistic license 2.0
Attribution Assurance Licenses
New and Simplified BSD licenses
Boost Software License (BSL1.0)
Computer Associates Trusted Open Source License 1.1
Common Development and Distribution License
Common Public Attribution License 1.0 (CPAL)
Common Public License 1.0
CUA Office Public License Version 1.0
EU DataGrid Software License
Eclipse Public License
Educational Community License, Version 2.0
Eiffel Forum License
Eiffel Forum License V2.0
Entessa Public License
Fair License
Frameworx License
GNU General Public License (GPL)
GNU General Public License version 3.0 (GPLv3)
GNU Library or "Lesser" General Public License (LGPL)
GNU Library or "Lesser" General Public License version 3.0 (LGPLv3)
Historical Permission Notice and Disclaimer
IBM Public License
Intel Open Source License
ISC License
Jabber Open Source License
Lucent Public License (Plan9)
Lucent Public License Version 1.02
Microsoft Public License (Ms-PL)

Microsoft Reciprocal License (Ms-RL)
MIT license
MITRE Collaborative Virtual Workspace License (CVW License)
Motosoto License
Mozilla Public License 1.0 (MPL)
Mozilla Public License 1.1 (MPL)
Multics License
NASA Open Source Agreement 1.3
NTP License
Naumen Public License
Nethack General Public License
Nokia Open Source License
Non-Profit Open Software License 3.0 (Non-Profit OSL 3.0)
OCLC Research Public License 2.0
Open Group Test Suite License
Open Software License 3.0 (OSL 3.0)
PHP License
Python license (CNRI Python License)
Python Software Foundation License
Qt Public License (QPL)
RealNetworks Public Source License V1.0
Reciprocal Public License
Reciprocal Public License 1.5 (RPL1.5)
Ricoh Source Code Public License
Simple Public License 2.0
Sleepycat License
Sun Industry Standards Source License (SISSL)
Sun Public License
Sybase Open Watcom Public License 1.0
University of Illinois/NCSA Open Source License
Vovida Software License v. 1.0
W3C License
wxWindows Library License
X.Net License
Zope Public License
zlib/libpng license

Nice to Know

FAIR - Licencing

Software vs. Data



About Open Source Licenses

Open source licenses are licenses that comply with the Open Source Definition — in brief, they allow software to be freely used, modified, and shared. To be approved by the Open Source Initiative (also known as the OSI), a license must go through the Open Source Initiative's license review process.

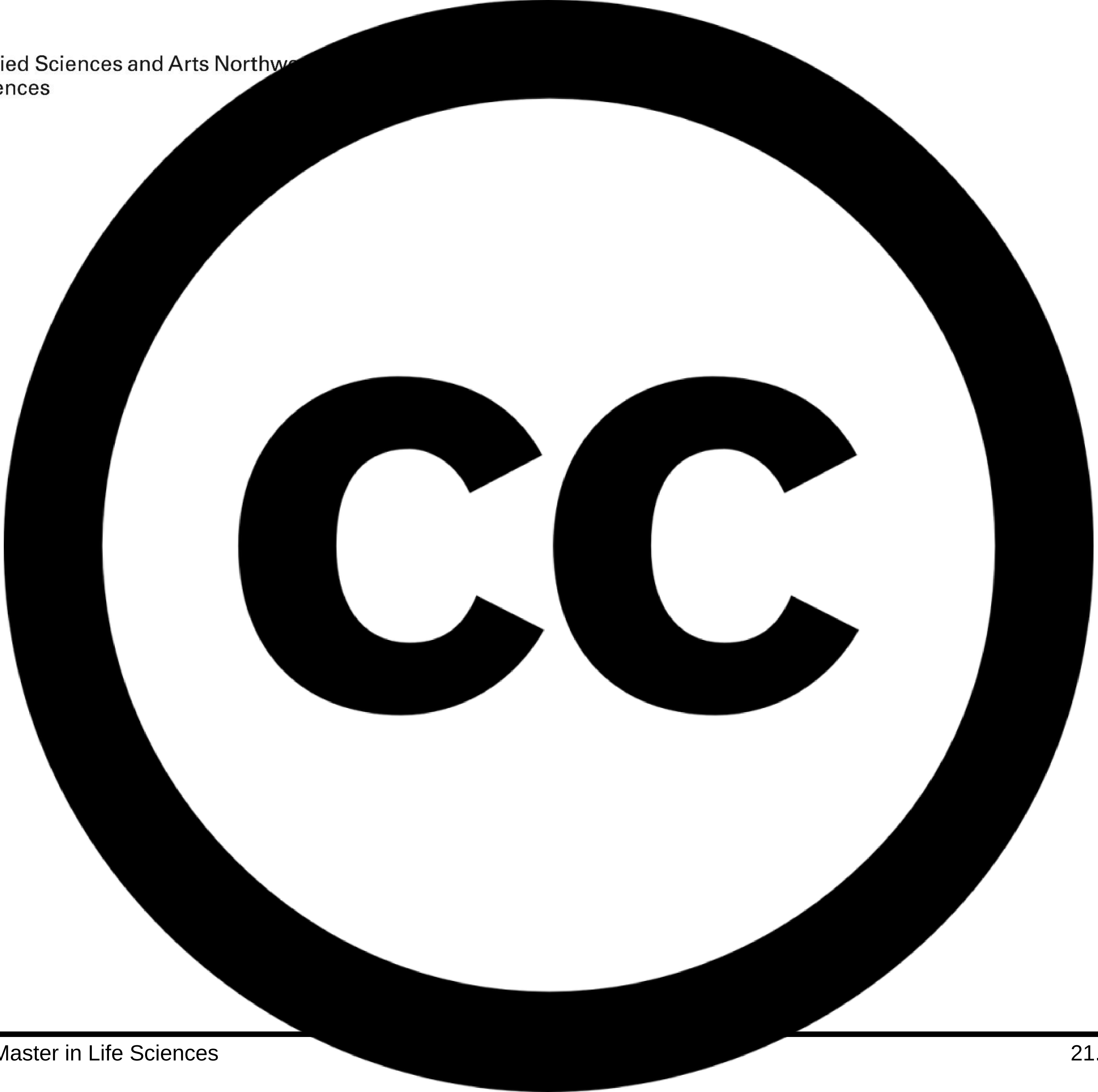
Popular Licenses

The following OSI-approved licenses are popular, widely used, or have strong communities:

- Apache License 2.0
- BSD 3-Clause "New" or "Revised" license
- BSD 2-Clause "Simplified" or "FreeBSD" license
- GNU General Public License (GPL)
- GNU Library or "Lesser" General Public License (LGPL)
- MIT license
- Mozilla Public License 2.0
- Common Development and Distribution License
- Eclipse Public License version 2.0

https://opensource.org/licenses

DATA



Nice to Know

FAIR - Licencing

Copyright for data sharing and fair dealing

When data are shared or archived, the original copyright owner retains the copyright.

A data archive cannot archive data unless all rights holders are identified and give their **permission** for the data to be shared. Secondary users need to obtain copyright clearance before data can be reproduced. However, exceptions exist under the fair dealing concept.

Nice to Know

FAIR - Licencing

Fair dealing

Under the fair dealing concept, data can be copied for non-commercial **teaching** or research purposes, private study, criticism or review without infringing copyright, provided that the owner of the work is sufficiently acknowledged.

This only applies to literary, dramatic, musical or artistic work, not to films or recordings.
An acknowledgement should give credit to the data source used, the data distributor and the copyright holder.

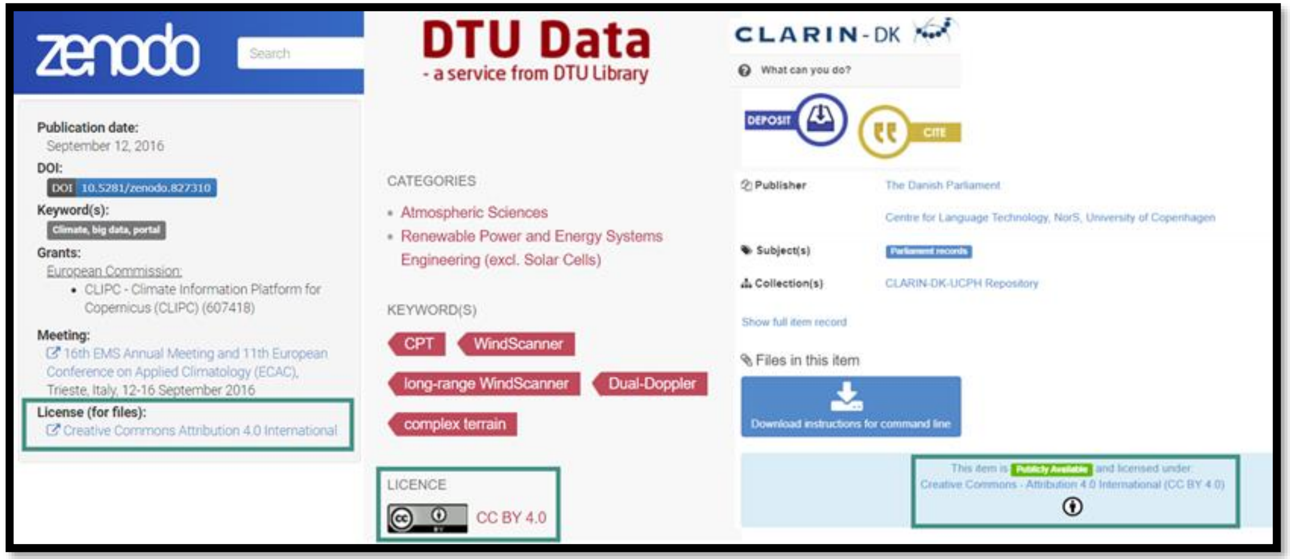
Nice to Know

FAIR - Licencing

What are data licences?

It’s great to publish - both data and articles - but releasing data without making the terms of use clear can be confusing and counterproductive, as it may deter potential users from using and citing the data. One of the most effective ways to communicate permissions to potential users of data are data licenses.

A data license is a legal arrangement between the creator of the data and the end-user, or the place the data will be deposited, specifying what users can do with the data.



Publication date: September 12, 2016

DOI: DOI 10.5281/zenodo.827310

Keyword(s): Climate, big data, portal

Grants: European Commission:

- CLIPC - Climate Information Platform for Copernicus (CLIPC) (607418)

Meeting: 16th EMS Annual Meeting and 11th European Conference on Applied Climatology (ECAC), Trieste, Italy, 12-16 September 2016

License (for files): Creative Commons Attribution 4.0 International

CATEGORIES

- Atmospheric Sciences
- Renewable Power and Energy Systems Engineering (excl. Solar Cells)

KEYWORD(S)

CPT WindScanner long-range WindScanner Dual-Doppler complex terrain

LICENCE

CC BY 4.0

Publisher	The Danish Parliament Centre for Language Technology, NorS, University of Copenhagen
Subject(s)	Parliament records
Collection(s)	CLARIN-DK-UCPH Repository

Show full item record

Files in this item

This item is Publicly Available and licensed under: Creative Commons - Attribution 4.0 International (CC BY 4.0)

https://www.howtofair.dk/how-to-fair/data-licences/

FAIR - Licencing

The CC licences

The most commonly and widely used data licences are the suite of **Creative Commons (CC)** copyright licences which clearly describe how data can and cannot be reused. The CC licences are irrevocable. This means that once you receive material under a CC licence, you will always have the right to use it under those licence terms, even if the licensor changes his or her mind and stops distributing under the CC licence terms.

Of course, you may choose to respect the licensor’s wishes and stop using the work, but once a dataset has been issued a CC licence, it cannot be revoked afterwards.

A scientific dataset, which other researchers may build upon or which is published together with a scientific article, is usually published under the CC-BY licence.

Nice to Know

FAIR - Licencing

Attribution CC BY

This licence lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licences offered. Recommended for maximum dissemination and use of licenced materials.

Nice to Know

FAIR - Licencing

Attribution CC BY-SA

This licence lets others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and licence their new creations under the identical terms. This licence is often compared to “copyleft” free and open source software licences. All new works based on yours will carry the same licence, so any derivatives will also allow commercial use. This is the licence used by Wikipedia, and is recommended for materials that would benefit from incorporating content from Wikipedia and similarly licenced projects.

Nice to Know

FAIR - Licencing

Attribution NoDerivs CC BY-ND

This licence lets others reuse the work for any purpose, including commercially. However, it cannot be shared with others in adapted form, and credit must be provided to you.

Nice to Know

FAIR - Licencing

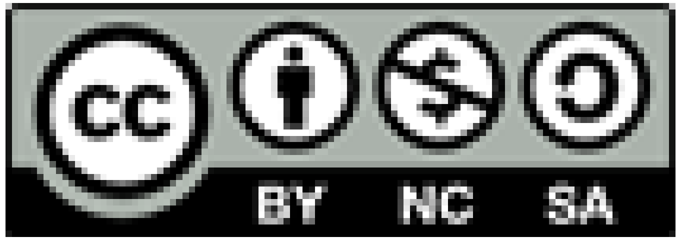
Attribution NonCommercial CC BY-NC

This licence lets others remix, tweak, and build upon your work non-commercially, and although their new works must also acknowledge you and be non-commercial, they don’t have to licence their derivative works on the same terms.

FAIR - Licencing

Attribution NonCommercial-ShareAlike CC BY-NC-SA

This licence lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and licence their new creations under the identical terms.

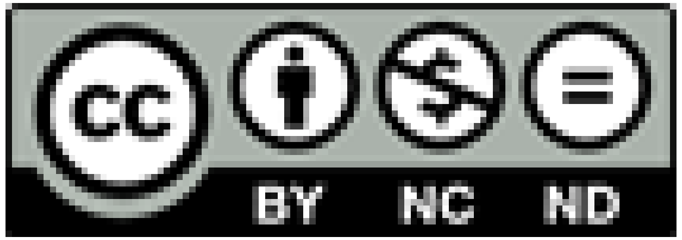


Nice to Know

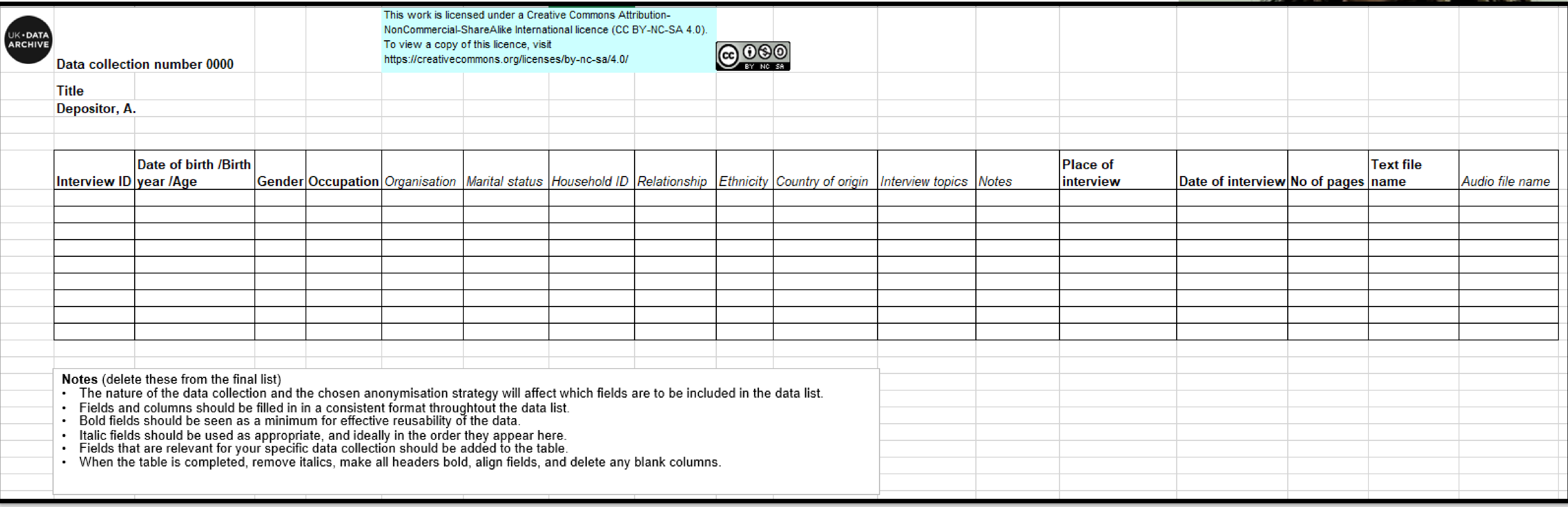
FAIR - Licencing

Attribution NonCommercial-NoDerivs CC BY-NC-ND

This licence is the most restrictive of these six main licences, only allowing others to download your works and share them with others as long as they credit you, but they can't change them in any way or use them commercially.



CC – Example (What does it mean??)



UK•DATA ARCHIVE

Data collection number 0000

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International licence (CC BY-NC-SA 4.0). To view a copy of this licence, visit https://creativecommons.org/licenses/by-nc-sa/4.0/

Title
Depositor, A.

Interview ID	**Date of birth /Birth year /Age**	**Gender**	**Occupation**	*Organisation*	*Marital status*	*Household ID*	*Relationship*	*Ethnicity*	*Country of origin*	*Interview topics*	*Notes*	**Place of interview**	**Date of interview**	**No of pages**	**Text file name**	*Audio file name*

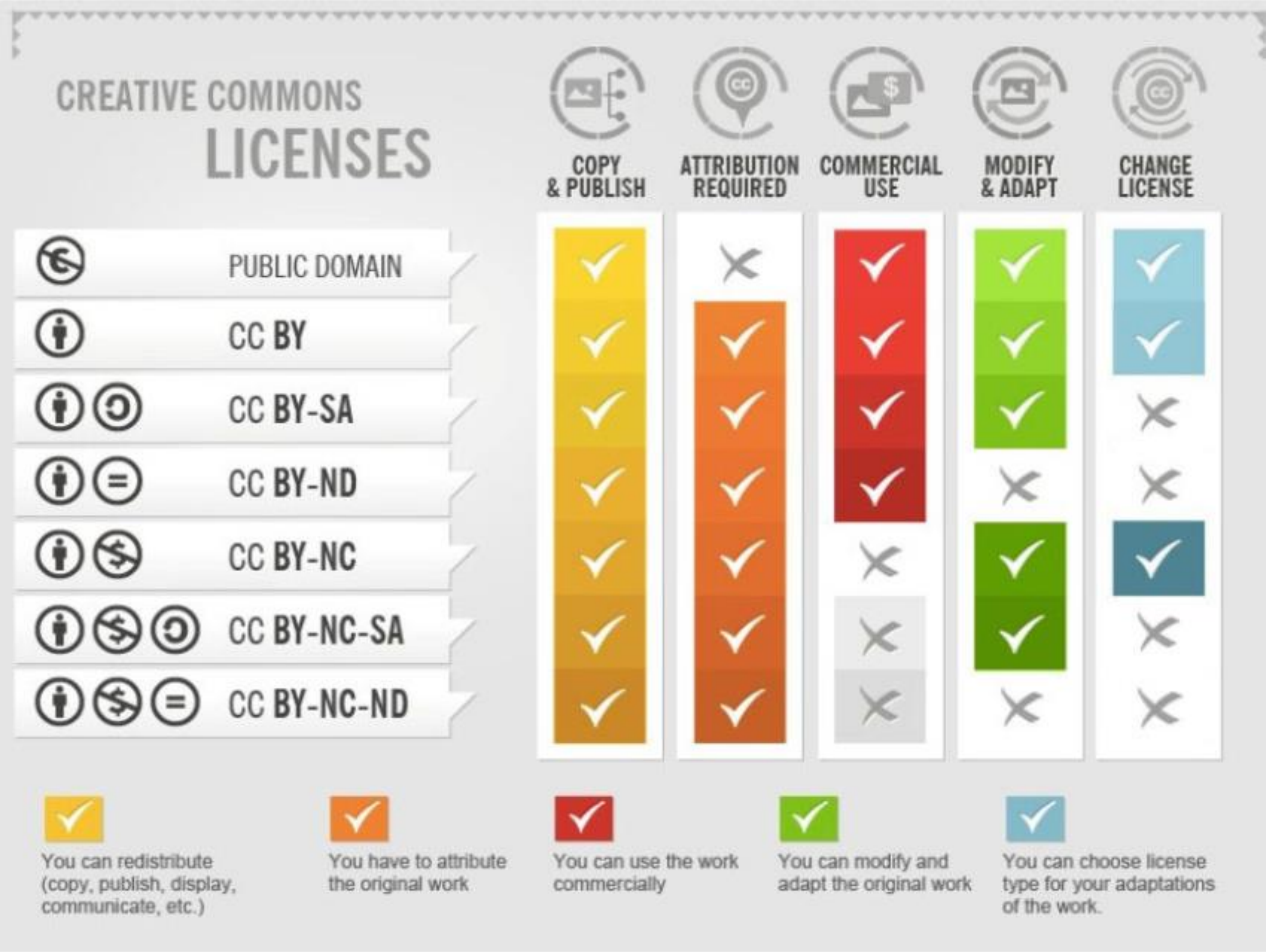
Notes (delete these from the final list)

- The nature of the data collection and the chosen anonymisation strategy will affect which fields are to be included in the data list.
- Fields and columns should be filled in in a consistent format throughtout the data list.
- Bold fields should be seen as a minimum for effective reusability of the data.
- Italic fields should be used as appropriate, and ideally in the order they appear here.
- Fields that are relevant for your specific data collection should be added to the table.
- When the table is completed, remove italics, make all headers bold, align fields, and delete any blank columns.

https://ukdataservice.ac.uk//app/uploads/uk_data_archive_data_listing_template.xlsx

Nice to Know

CC - Summary



CREATIVE COMMONS LICENSES

	COPY & PUBLISH	ATTRIBUTION REQUIRED	COMMERCIAL USE	MODIFY & ADAPT	CHANGE LICENSE
PUBLIC DOMAIN	✓	✗	✓	✓	✓
CC BY	✓	✓	✓	✓	✓
CC BY-SA	✓	✓	✓	✓	✗
CC BY-ND	✓	✓	✓	✗	✗
CC BY-NC	✓	✓	✗	✓	✓
CC BY-NC-SA	✓	✓	✗	✓	✗
CC BY-NC-ND	✓	✓	✗	✗	✗

- ✓ You can redistribute (copy, publish, display, communicate, etc.)
- ✓ You have to attribute the original work
- ✓ You can use the work commercially
- ✓ You can modify and adapt the original work
- ✓ You can choose license type for your adaptations of the work.

https://foter.com/blog/how-to-attribute-creative-commons-photos/

FAIR – Data Level Documentation

Data-level documentation

Data-level documentation provides information about individual databases or data files. This could be, for example, interview transcripts or pictures, as well as documentation for elements within the files, for example describing variables within an SPSS file.

Upcoming….

1. Qualitative data
2. Quantitative data
3. Secondary sources

FAIR – Qualitative Data

Qualitative data

For qualitative textual data, the background, contextual information, participant details of interviews, observations or diaries, can all be described at the beginning of a file as a header or summary page.

Clear speech demarcation and the use of speaker tags are crucial in interview transcripts. Examples can be seen in our model transcription template.

FAIR – Qualitative Data

Qualitative data

For qualitative data collections, such as interview or image collections, an important piece of data documentation is the **data list**, which accompanies the data collection in our catalogue.

The list provides information for users that enables them to easily identify and locate relevant transcripts or items within a data collection. Each item in the list should have a unique identifier. The list provides key biographical characteristics and features of interviewees, plus details for the interview, for example:

interview ID, age, gender, occupation, organization, location, place of interview, date of interview, transcript file name, recording file name

FAIR – Qualitative Data

Quantiative data

With quantitative data, data documentation can be embedded within data files, such as variable and code descriptions in databases.

Many data analysis software packages have facilities for data annotation and description, as variable attributes, data type definitions, table relationships and so on. Alternatively, information about data items can be recorded in a structured document such as a codebook

Structured tabular data should have as documentation (where applicable):

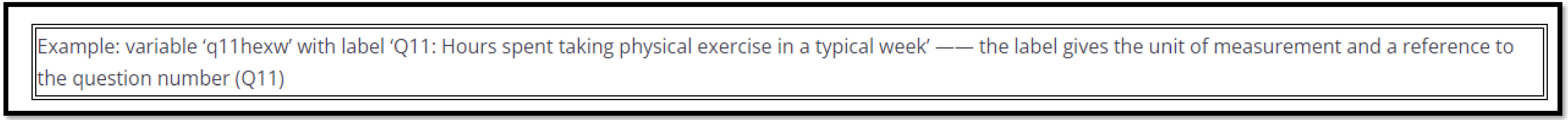
- Variable names, labels and descriptions (maximum 80 characters).
- Units of measurement for variables.
- Reference to the question number of a survey or questionnaire.

FAIR – Qualitative Data

Quantiative data

Structured tabular data should have as documentation (where applicable):

- Variable names, labels and descriptions (maximum 80 characters).
- Units of measurement for variables.
- Reference to the question number of a survey or questionnaire.



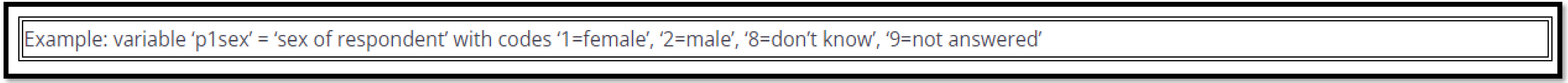
Example: variable ‘q11hexw’ with label ‘Q11: Hours spent taking physical exercise in a typical week’ —— the label gives the unit of measurement and a reference to the question number (Q11)

How FAIR?

FAIR – Qualitative Data

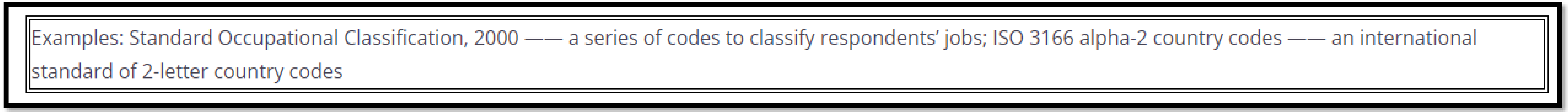
Quantiative data

- Value code labels



Example: variable ‘p1sex’ = ‘sex of respondent’ with codes ‘1=female’, ‘2=male’, ‘8=don’t know’, ‘9=not answered’

- Coding and classification schemes explained, with a bibliographic and dated reference (some standards change over time).

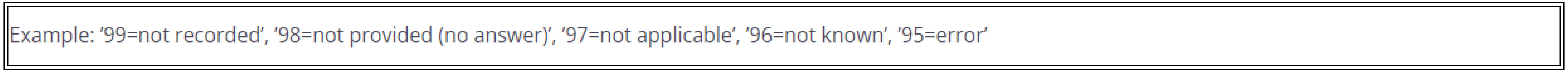


Examples: Standard Occupational Classification, 2000 —— a series of codes to classify respondents’ jobs; ISO 3166 alpha-2 country codes —— an international standard of 2-letter country codes

FAIR – Qualitative Data

Quantiative data

- Codes for missing data, with reason data are missing (blanks, system-missing or ‘0’ values are best avoided).

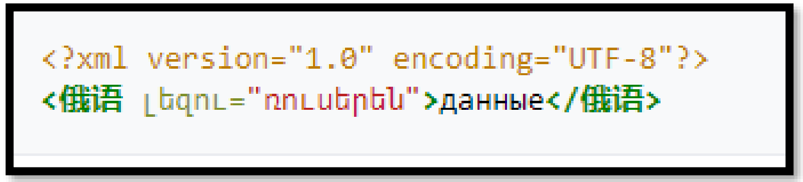


Example: '99=not recorded', '98=not provided (no answer)', '97=not applicable', '96=not known', '95=error'

- Defining placeholder for variables in case of skipped cases or questions.
- Derived or constructed variables created after collection, giving code, algorithm or command files used to create them —— simple derivations, such as grouping age data into age intervals, can be explained in the variable and value labels; complex derivations can be described by providing the algorithms, logical statements or functions used to create derived variables, such as the SPSS or Stata command files. and classification schemes explained, with a bibliographic and dated reference (some standards change over time).

How FAIR?

FAIR – Qualitative Data



```
<?xml version="1.0" encoding="UTF-8"?>
<俄语 լեզու="ռուսերեն">данные</俄语>
```

Quantiative data

Many data software packages have facilities for data annotation and description as variable attributes (labels, codes, data type, missing values), table relationships, etc..

Example embedded documentation SPSS file: Variable descriptions and attributes, such as codes, data type, missing values, can be documented for each variable in ‘Variable View’ or via syntax, whereby embedded data documentation is then contained in the SPSS command file.

Example embedded documentation MS Access database: Variable descriptions and attributes can be documented in ‘Design View’ and relationships between tables and files can be created.

GIS e.g ArcGIS: Shapefiles or layers and tables can be organised in a geo-database with rich metadata created in ArcCatalog.

Example embedded documentation MS Excel spreadsheet: An additional worksheet within the data file can contain variable and data-related documentation.

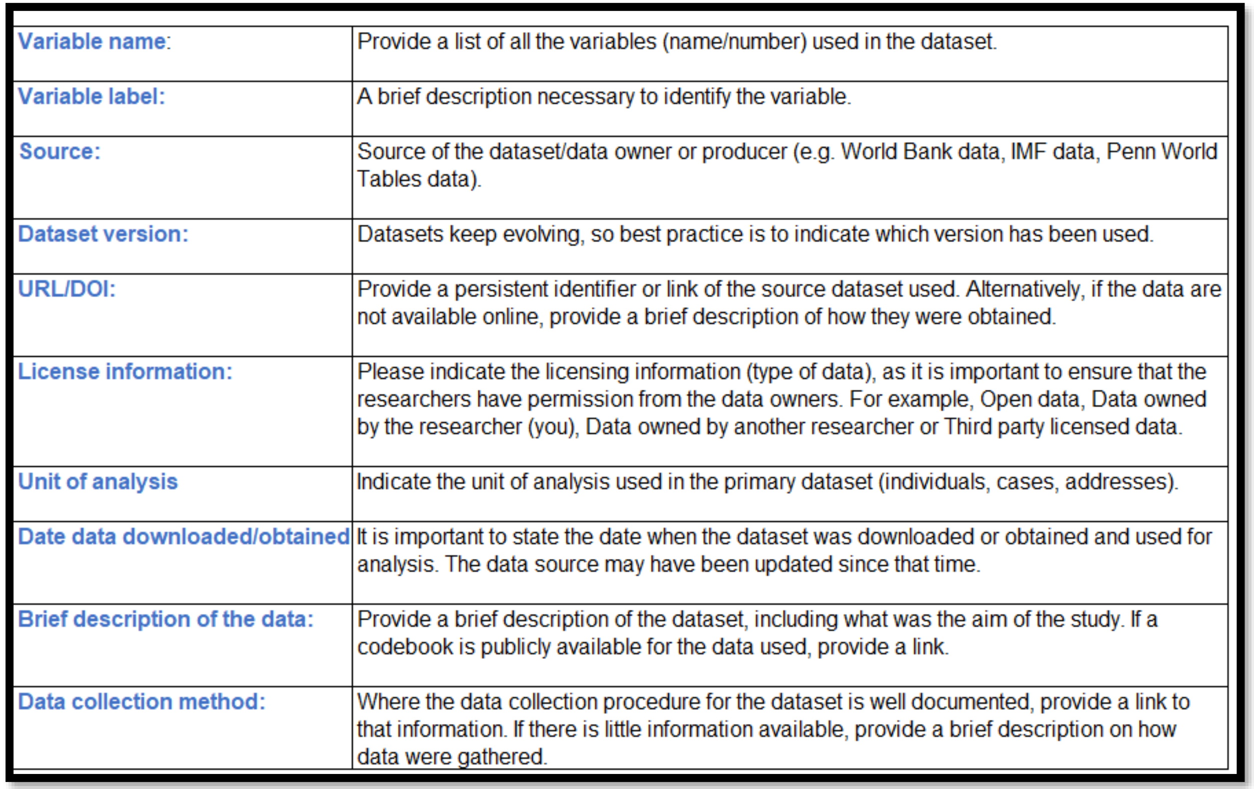
How FAIR?

FAIR – Secondary Data

For datasets being deposited that include secondary data resources, researchers are advised to prepare a variable information log describing these resources.

An example is where a primary data source used may have particular restrictions placed upon its use and any subsequent use. Data gathered from a website may seem like it is ‘open’, but it may come with limitations on processing, publishing and further dissemination.

variable information log template = the information required to ensure that researchers are able to clearly understand secondary data.



Variable name:	Provide a list of all the variables (name/number) used in the dataset.
Variable label:	A brief description necessary to identify the variable.
Source:	Source of the dataset/data owner or producer (e.g. World Bank data, IMF data, Penn World Tables data).
Dataset version:	Datasets keep evolving, so best practice is to indicate which version has been used.
URL/DOI:	Provide a persistent identifier or link of the source dataset used. Alternatively, if the data are not available online, provide a brief description of how they were obtained.
License information:	Please indicate the licensing information (type of data), as it is important to ensure that the researchers have permission from the data owners. For example, Open data, Data owned by the researcher (you), Data owned by another researcher or Third party licensed data.
Unit of analysis	Indicate the unit of analysis used in the primary dataset (individuals, cases, addresses).
Date data downloaded/obtained	It is important to state the date when the dataset was downloaded or obtained and used for analysis. The data source may have been updated since that time.
Brief description of the data:	Provide a brief description of the dataset, including what was the aim of the study. If a codebook is publicly available for the data used, provide a link.
Data collection method:	Where the data collection procedure for the dataset is well documented, provide a link to that information. If there is little information available, provide a brief description on how data were gathered.

How FAIR?

Codebook

A *codebook* is a document containing information about each of the variables in your dataset, such as:

- The name assigned to the variable
- What the variable represents (i.e., its label)
- How the variable was measured (e.g. nominal, ordinal, scale)
- How the variable was actually recorded in the raw data (i.e. numeric, string; how many characters wide it is; how many decimal places it has)
- For scale variables: The variable's units of measurement
- For categorical variables: If coded numerically, the numeric codes and what they represent

Heavy example:
https://ddialliance.org/sites/default/files/National%20Household%20Survey,%202011%20%5bCanada%5d%20Public%20Use%20Microdata%20File%20(PUMF)-%20Individuals%20File.pdf

Light Example:
https://gist.github.com/JorisSchut/dbc1fc0402f28cad9b41

File Naming

Descriptive file names are an important part of organizing, sharing, and keeping track of data files. Develop a naming convention based on elements that are important to the project.

File naming best practices:

- ❑ Files should be named consistently
- ❑ File names should be short but descriptive (<25 characters) (Briney, 2015)
- ❑ Avoid special characters or spaces in a file name
- ❑ Use capitals and underscores instead of periods or spaces or slashes
- ❑ Use date format ISO 8601: YYYYMMDD
- ❑ Include a version number (Creamer et al. 2014)
- ❑ Write down naming convention in data management plan

Briney, K. (2015) Data management for researchers : organize, maintain and share your data for research success. Exeter, UK: Pelagic Publishing.

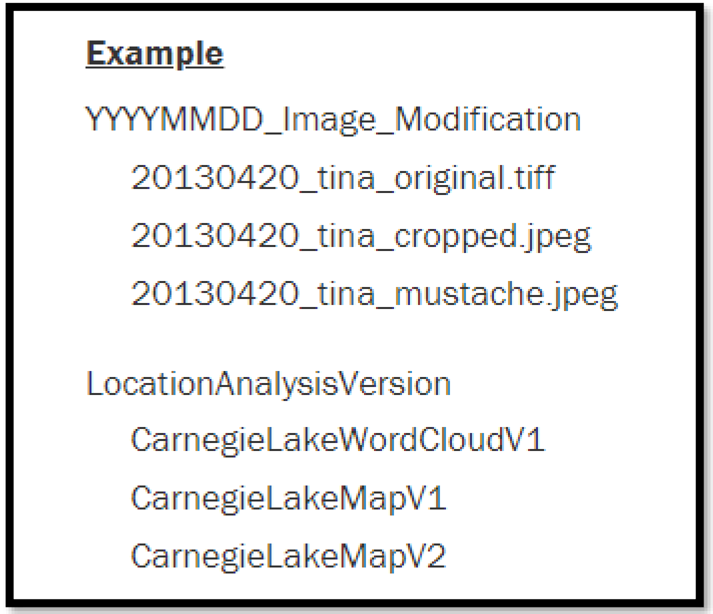
Creamer AT, Martin ER, Kafel D. (2014). Research Data Management and the Health Sciences Librarian. Library Publications and Presentations. Retrieved from https://escholarship.umassmed.edu/lib_articles/147

How FAIR?

File Naming

Elements to consider using in a naming convention are:

- ❑ Date of creation (putting the date in the front will facilitate computer aided date sorting)
- ❑ Short Description
- ❑ Work
- ❑ Location
- ❑ Project name or number
- ❑ Sample
- ❑ Analysis
- ❑ Version number



Example

YYYYMMDD_Image_Modification

20130420_tina_original.tiff

20130420_tina_cropped.jpeg

20130420_tina_mustache.jpeg

LocationAnalysisVersion

CarnegieLakeWordCloudV1

CarnegieLakeMapV1

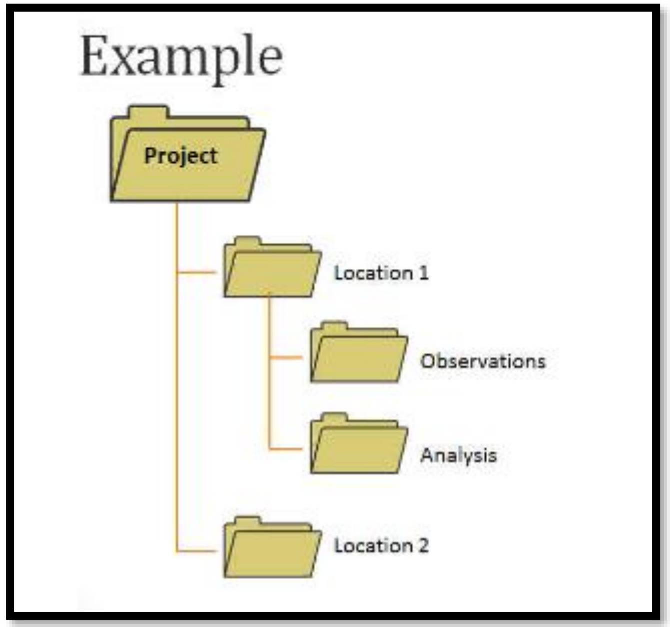
CarnegieLakeMapV2

How FAIR?

File Structure

Hierarchical file structures can add additional organization to your files. As with file naming use whatever makes most sense for your data. Some possibilities include:

- ❑ Project
- ❑ Date
- ❑ Analysis
- ❑ Location

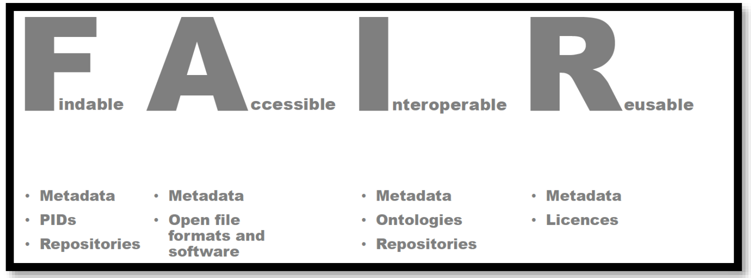


Outlook

Coaching Session 5

Take your project described in your DMP.

1. Prepare a folder structure (a model)
2. Suggest a file naming convention (a template)
3. Create a dummy dataset (e.g. in Excel)
4. Create a dummy codebook for this dataset (e.g., in Excel)
5. Think of FAIR requirements relevant for this dataset:



6. Describe this process on the example of this dataset in your handbook