Getting started with Metadata Breakout Room
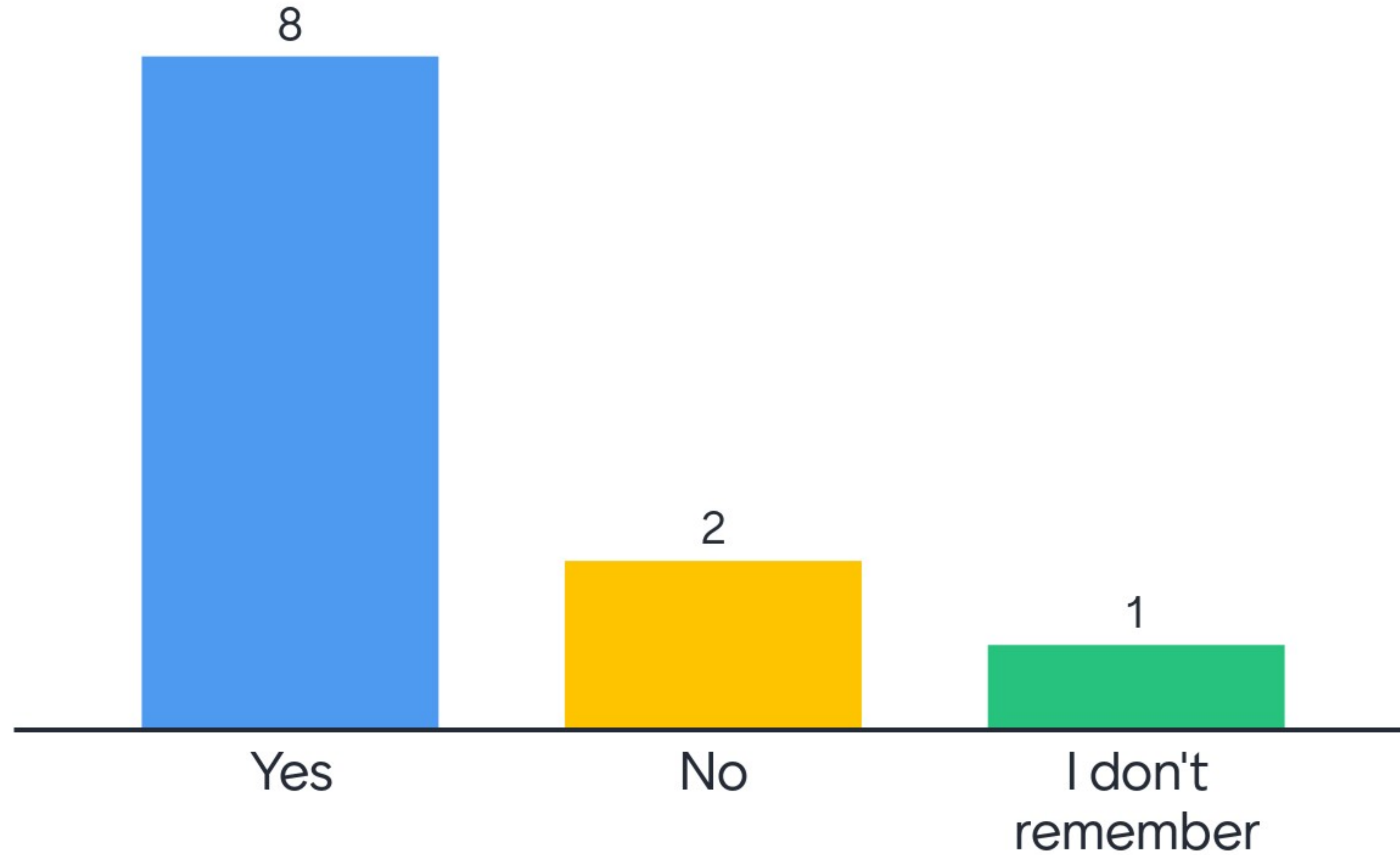
Kayla Burnim from RITMO

Breakout Room Overview

→ An engineer's systematic approach to getting started with metadata

→ Interactive presentation using Menti

→ Break

→ Question and answer

Have you used Menti before?

What are your goals for this breakout session?

Going a bit more details. Examples

learning your experience with metadata and management

Practical examples of metadata storage
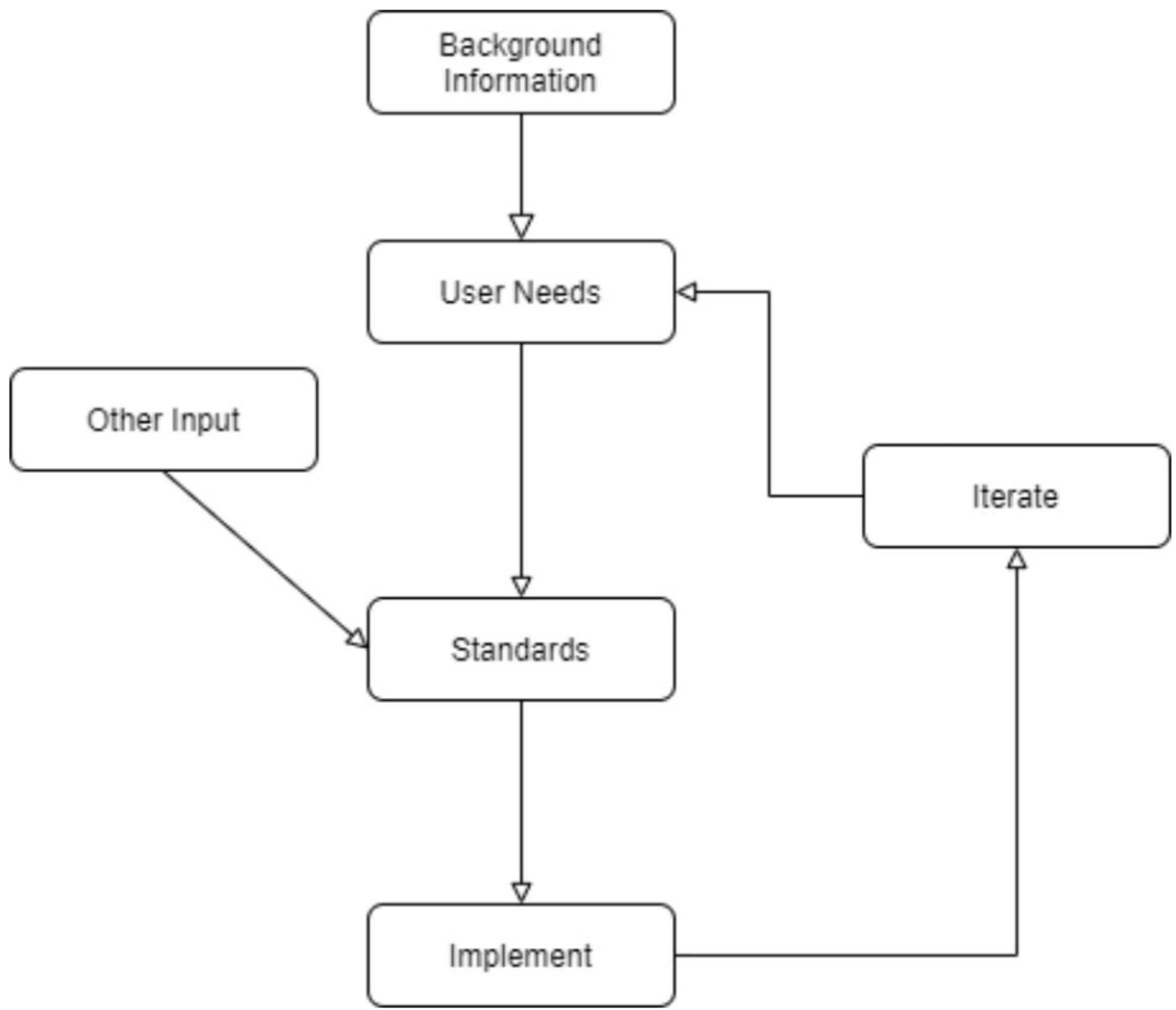
Lean more about metadata

understanding the role of metadata in replication

get to know the basics of metadata (where to start etc)

Learn about your project. Get inspired on how to advice our researchers on data management

Tips for appropriate software for metadata storage

More information about managing metadata for diverse projects.

Systemic Approach to Metadata

→ Background

→ User Needs

→ Other Inputs

→ Standards

→ Implementation

Background

Before diving into user needs, it is important to collect relevant background information.
- User information
- Research methods
- IT infrastructure
- Data management
- Current metadata

RITMO Background Example

→ Psychology (Social Sciences) + Informatics (Mathematics and Natural Sciences) + Musicology (Humanities)

→ Researchers are affiliated with one (or more) of the three departments

→ Large variety of computer skills

→ Over 60 people employed at RITMO

→ Qualitative methods from the arts and humanities such as interviews, qualitative analyses, and aesthetic and cultural interpretation

→ Quantitative methods from psychology and neuroscience such as behavioral experiments, motion capture and neuroimaging techniques

→ Informatics, including machine learning, computer modelling, and rapid prototyping

→ Research methods are consistently inconsistent

→ Data Management was unique to the individual

→ Data is saved in many locations

→ Most data was undocumented

→ Researchers were not confident with data management

What is some of your background?

data manager

material engineer

Subject librarian/research support (with little knowledge on how to choose metadata)

humanities, work with human participants, little IT knowledge at the faculty

Colour Imaging research and teaching. Appearance measurement, Psychophysical experiments

computer science

IT Developer

Research support, research experience from the humanities

Research advisor in Humanities

15

What is some of your background?

researcher, humanities	geosciences	Animal science researcher

library and information scienceresearch supportlanguages	linguistics	not sure: access rights, title, data owner, description, date

15

Generating user needs

What do the researchers generating the data need from the metadata to make their data understandable to others?

Brainstorm as big of a list as possible. Apply appropriate background information to generate this list.

Final list can be broken into three categories: must have (needed for minimal viable product), nice to have (important but not a dealbreaker), and "bonus" wants.

	A	B
1	**Element**	
2	Title	Must
3	Description	Nice
4	PI/Data owner	Must
5	Other people involved	Nice
6	Institution	Bonus
7	Funder	Bonus
8	Related Projects	Bonus
9	Linked publications	Bonus
10	Data Type	Nice
11	File size	Bonus
12	Equipment used	Nice
13	Date of creation	Nice
14	Privacy approval	Nice
15	Copyright information	Nice
16	Other legal/ethical	Bonus
17	**Functionality**	
18	Quick to complete	Nice
19	Minimal coding needed	Must
20	Editable	Must
21	Transitions to perminent solution	Must
22	Multidisciplinary	Must
23	Machine Readable	Nice
24	Automatic Tool	Bonus

User needs example

Example list of user needs from RITMO example. Includes must have, nice to have, and bonus categories. Additionally, needs divided into elements and functionality

User Needs: What are some of your must haves?

title - data owner - equipment used - user license

name, IP owner, institution, title,

source of the primary data, date of collection, secondary source (if any)

Basic metadata to enabling retrieval

title, owner, produced date(s)

Measurement details that identify trace them for easy repeat

description, source of data, how to use it, date

and keywords for retrieval

not sure: access rights, title, data owner, data collection date, text type

11

User Needs: What are some of your must haves?

Details of who performed measurements. or where

publication, timeperiode covered

User Needs: What are some of your nice to have?

funder - related publications - geographic location - time period covered

production date

completeness of a standard

Who performed the measurements or where.

filetype, last date changed, time span

data collectors, funding information

6

Other Inputs

→ Current or future institution requirements or recommendations

→ Repository requirements or guidelines

→ Funder or other body have requirements, guidelines, recommendations, etc

→ Target publication journal requirements

→ Is this a temporary or permanent metadata solution?

→ Is this for a single project or system for multiple projects?

Examples of other inputs

At RITMO

- UiO has metadata system in development
- Need temporary, flexible solution
- Other lab group using BIDS
- No single journal target
- Archives vary but requirements general

Other Inputs: What are some other inputs that you have?

user requrements

archive standards

recommended citation

data management plan

funding requirements

enabling citations

embargo maybe

What is a Metadata Standard?

A Standard provides a structure to describe data with:

- Common terms to allow consistency between records
- Common definitions for easier interpretation
- Common language for ease of communication
- Common structure to quickly locate information

In search and retrieval, standards provide:

- Documentation structure in a reliable and predictable format for computer interpretation
- A uniform summary description of the dataset

https://dataoneorg.github.io/Education/lessons/07_metadata/07_metadata.pdf

Finding Standards

The "just google it and figure it out" procedure can be overwhelming and time consuming due to not only the quantity of metadata standards but also the resources available to create them.

In this section:
- Where to find standards
- 3 methods for narrowing down your choices

Where to find standards

https://rd-alliance.github.io/metadata-directory/standards/

https://www.dcc.ac.uk/guidance/standards/metadata/list

https://fairsharing.org/standards/

If "Other Inputs" dictate metadata

In cases that your institution, funder, repository
etc require specific metadata standard, use that!

But wait, those "user needs"
Check the user needs against the other input
documentation. Is it missing anything? In some
cases you may make an "optional" item required
while in others additional metadata might be
needed.

Customize XML Example https://www.ebi.ac.uk/ena/browser/view/ERC000015
https://www.ebi.ac.uk/ena/browser/api/xml/ERC000015

Subject Specific Standards

Some research subjects and data types have
specific standards. Many are governed by ISO
standards.
Benefits include subject specific elements and
increased ease of sharing. Some also have tools
to read expected files, generate folder structures,
and other data management.
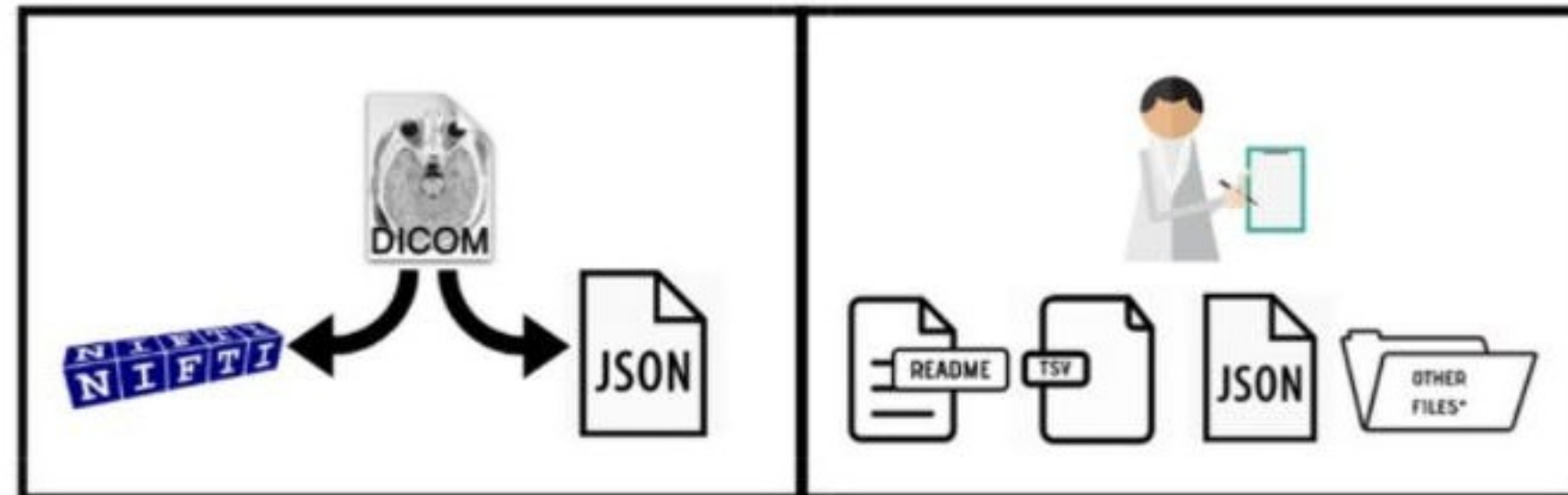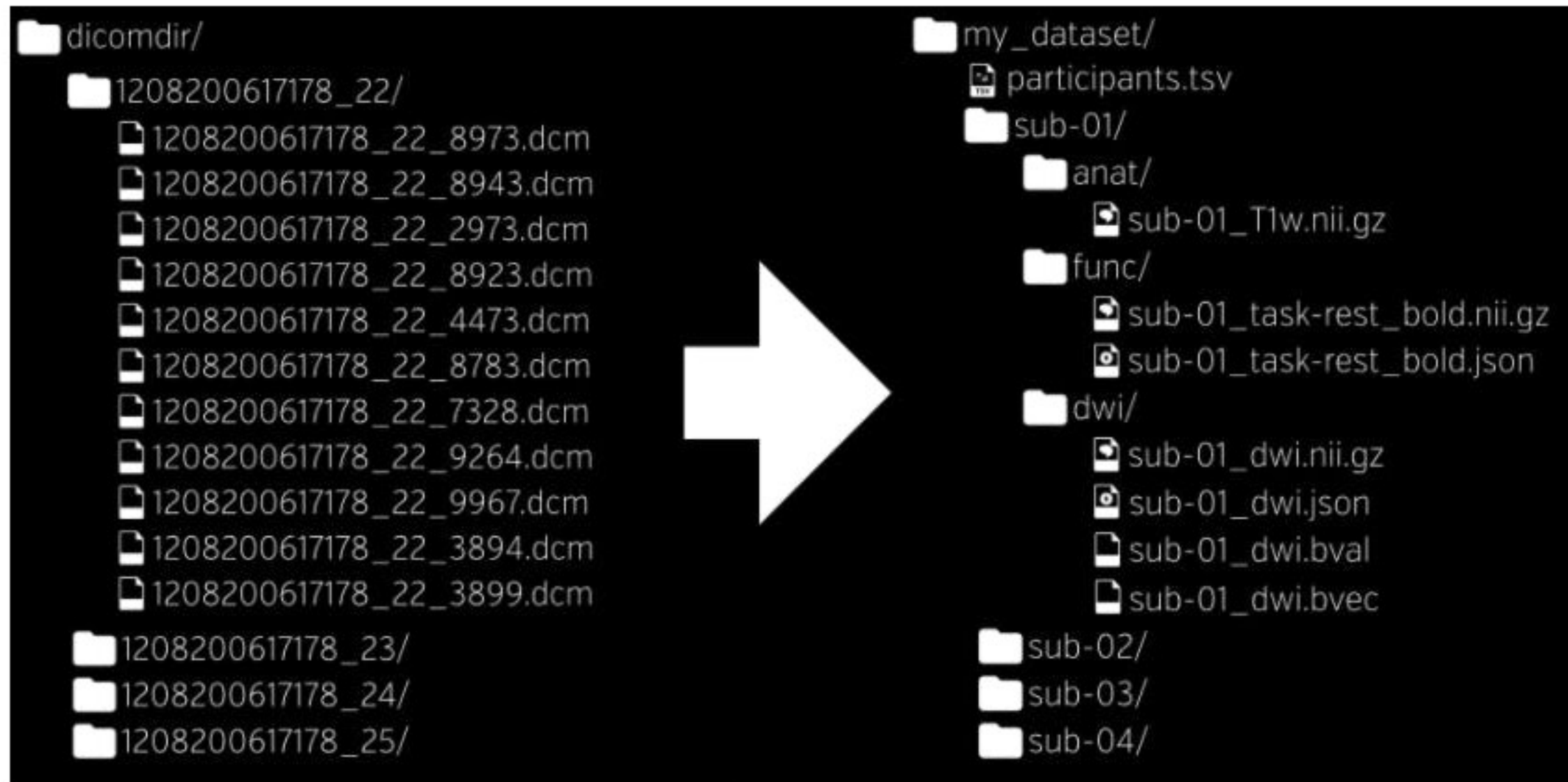
Examples of subject specific standards

BIDS (Brain Imaging Data Structure) data standard developed for fMRI. Tools generate:
- README file with basic information, like study name, data's owners, collaborators, location of the data and study protocols.
- Structured file naming convention with participant, session, type of data (eg. structural MR scan, fMRI or EEG) and task a file pertains to.
- JSON machine readable metadata standard

https://www.sv.uio.no/psi/english/research/projects/human-time-data/documents/BIDS/
https://bids.neuroimaging.io/

General Standards

In some cases, no subject specific standard will be applicable. A variety of general standards exist.

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional given name, family name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

Table 2: DataCite Recommended and Optional Properties

ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier, and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box, place, and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O
20	RelatedItem (with identifier, creator, title, publication year, volume, issue, number, page, publisher, edition, and contributor sub-properties)	O

DC Project Metadata Template.txt - Notepad

File   Edit   Format   View   Help

```
1. TITLE  (Your Project Name)
The name given to the resource
2. CREATOR (Project Lead)
The person primarily responsible for the intellectual content of the resource; the author.
3. SUBJECT (Main Topic)
The topic of the resource written in keyword format
4. DESCRIPTION
A textual description of the content of the resource
5. PUBLISHER (Probably UiO)
The entity responsible for making the resource available
6. CONTRIBUTORS
Additional person(s) or organization(s) responsible for making contributions to the project
7. DATE (Beginning of project)
The date associated with the beginning of the project.
8. TYPE
The nature or genre of the content of the resource.
9. FORMAT
The physical or digital format of the resource
10. IDENTIFIER (Optional)
A string or number used to uniquely identify the resource.
11. SOURCE (Optional)
The reference from which the resource is delivered (if applicable).
12. LANGUAGE
The language(s) of the intellectual content of the resource.
13. RELATION (Links to other projects)
A reference to a related resource
14. COVERAGE
The spatial and/or temporal scope of the resource
15. RIGHTS MANAGEMENT (If applicable)
Information about rights held in and over the resource.
```

HTTP://SCHEMA.DATACITE.ORG/META/KERNEL-4.4/DOC/DATACITE-METADATAKERNEL_V4.4.PDF

Examples of General Standards - DataCite and DublinCore https://www.dublincore.org/

What type of standard do/will you need?
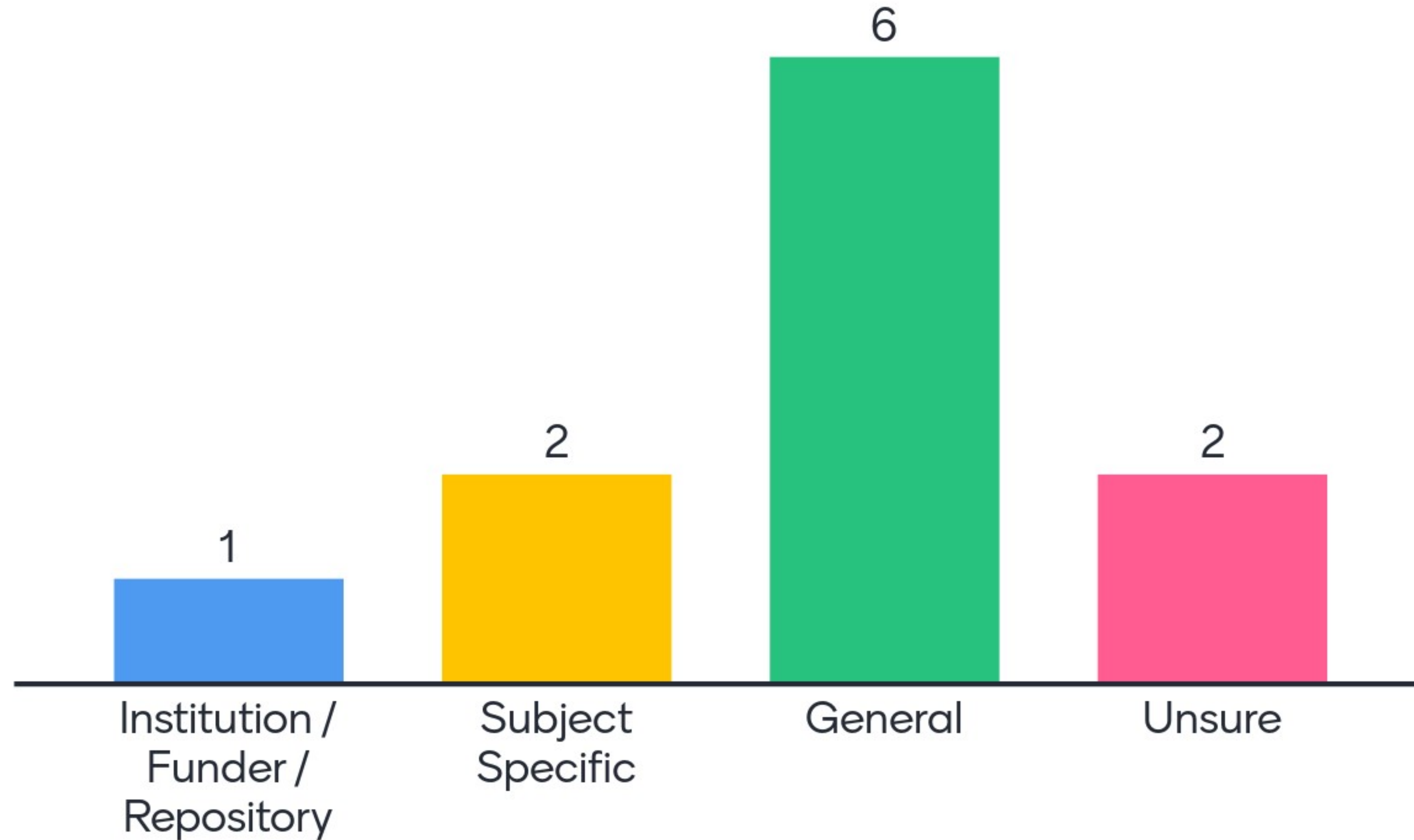
Matching User Needs with Standards

Depending on your user needs, standard type,
preference, etc the method for matching might
differ. Terminology in the elements of a standard
might vary as well so keep that in mind.
- Simple example matching with a single standard
- Systematic decision making method

Simple case example

	User Needs		Dublin Core		DataCite	
Element						
Title	Must	10	Yes	10	Mandatory	10
Description	Nice	5	Yes	5	Recommended	5
PI/Data owner	Must	10	Yes	10	Mandatory	10
Other people involved	Nice	5	Yes	5	Recommended	5
Institution	Bonus	1	Yes	1	No	0
Funder	Bonus	1	No	0	Optional	1
Related Projects	Bonus	1	Yes	1	Recommended	1
Linked publications	Bonus	1	No	0	No	0
Data Type	Nice	5	Yes	5	Mandatory	5
File size	Bonus	1	No	0	Optional	1
Equipment used	Nice	5	No	0	No	0
Date of creation	Nice	5	Yes	5	Recommended	5
Privacy approval	Nice	5	No	0	No	0
Copyright information	Nice	5	Yes	5	Optional	5
Other legal/ethical	Bonus	1	No	0	No	0
Functionality						
Quick to complete	Nice	5	Yes	5	Yes	5
Minimal coding needed	Must	10	Yes	10	Yes	10
Editable	Must	10	Yes	10	Yes	10
Transitions to perminent solution	Must	10	Yes	10	Yes	10
Multidisciplinary	Must	10	Yes	10	Yes	10
Machine Readable	Nice	5	Optional	5	Optional	5
Automatic Tool	Bonus	1	Optional	1	Unknown	0
	Sum	112		98		98

Systematic matching

Note on Machine Readablility

Metadata facilitate cataloguing data and data
discovery. Metadata are intended for machine-reading
and makes your data more FAIR . It is considered best
practice to have the metadata machine readable
format (i.e., XML, JSON, CSV, RDF, HTML).
Some of you may use temporary metadata solutions
but the final format goal should be machine readable

Metadata Tools

There are a variety of metadata generating
tools available for some standards.

Main Page   **Simple Generator**   **Advanced Generator**   **xZINECOREx Generator**   **About**   **Contribute**

Main Page

Welcome to dublincoregenerator.com! Our mission is to provide accurate and dynamic tools for generating Dublin Core metadata code.

If you want to generate simple Dublin Core metadata using only the 15 main elements, use the Simple Generator. If you are interested in generating more advanced Dublin Core metadata code using the more detailed qualified elements and encoding schemes, use the Advanced Generator. There's also a generator for the xZINECOREx variation of Dublin Core.

Feel free to also read up about the site and how you can help.

Final Needs Input

→ Does the standard chosen have the appropriate level of machine readability?

→ Are there tools for the standard?

→ What is the output format?

→ Need a different standard?

Make your own

Cedar:
https://metadatacenter.org/
FAIR Cookbook:
https://fairplus.github.io/the-fair-cookbook/content/recipes/interoperability/creating-minimal-metadata-profiles.html

Implementation

Now that a standard and method for generating metadata has been chosen, additional pieces of implementation can be added

Answering the question "how do I complete the metadata?"
Includes instructional materials, training, examples, help/support, and even vocabulary

Standards have instructions

https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

http://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf

Table 3: Expanded DataCite Mandatory Properties

ID	DataCite-Property	Occ	Definition	Allowed values, examples, other constraints
1	Identifier	1	The Identifier is a unique string that identifies a resource. For software, determine whether the identifier is for a specific version of a piece of software, (per the Force11 Software Citation Principles[11]), or for all versions.	DOI (Digital Object Identifier) registered by a DataCite member. Format should be "10.1234/foo"
1.a	identifierType	1	The type of Identifier.	Controlled List Value: DOI
2	Creator	1-n	The main researchers involved in producing the data, or the authors of the publication, in priority order. To supply multiple creators, repeat this property.	May be a corporate/institutional or personal name. Note: DataCite infrastructure supports up to 10,000 names. For name lists above that size, consider attribution via linking to the related metadata.
2.1	creatorName	1	The full name of the creator.	Examples: Charpy, Antoine; Jemison, Mae; Foo Data Center Note that the personal name, format should be: family, given. Names in non-roman scripts may be transliterated according to the ALA-LC tables[12].
2.1.a	nameType	0-1	The type of name	Controlled List Values: Organizational Personal (default)

Term Name: creator | More details

URI	http://purl.org/dc/terms/creator
Label	Creator
Definition	An entity responsible for making the resource.
Comment	Recommended practice is to identify the creator with a URI. If this is not possible or feasible, a literal value that identifies the creator may be provided.
Type of Term	Property
Range Includes	• http://purl.org/dc/terms/Agent
Equivalent Property	• http://xmlns.com/foaf/0.1/maker
Subproperty of	• Creator (http://purl.org/dc/elements/1.1/creator) • Contributor (http://purl.org/dc/terms/contributor)

Term Name: date | More details

URI	http://purl.org/dc/terms/date
Label	Date
Definition	A point or period of time associated with an event in the lifecycle of the resource.
Comment	Date may be used to express temporal information at any level of granularity. Recommended practice is to express the date, date/time, or period of time according to ISO 8601-1 [ISO 8601-1] or a published profile of the ISO standard, such as the W3C Note on Date and Time Formats [W3CDTF] or the Extended Date/Time Format Specification [EDTF]. If the full date is unknown, month and year (YYYY-MM) or just year (YYYY) may be used. Date ranges may be specified using ISO 8601 period of time specification in which start and end dates are separated by a '/' (slash) character. Either the start or end date may be missing.
Type of Term	Property
Has Range	http://www.w3.org/2000/01/rdf-schema#Literal
Subproperty of	• Date (http://purl.org/dc/elements/1.1/date)

UiO : RITMO
University of Oslo

RITMO Metadata Help

→ Internal Handbook with instructions

→ Examples, templates, and links

→ One-on-One meetings

→ Group check-in during lab meetings

→ Data Management Office Hours

→ Training and seminar opportunities

Sharing Best Practices

→ Make the metadata as useful as possible!

→ Standardized Vocabulary

→ https://lov.linkeddata.es/dataset/lov/

→ Example Dates: 29May2021, 29052021, May 29th 2021, 20210521

"Babysitting" phase

Data management gets bumped for other
deadlines
It can be new, scary, and overwhelming
"I'll do it later" turns into "I forgot about it"

Creating many chances for teamwork, having
one on one meetings, scheduling office hours,
checking in during lab meetings, etc can increase
completion percentages

Assessment and Iteration

→ Is the standard capturing what you need?

→ Did you learn of a new system to try?

→ What is the level of completion and accuracy?

→ Where is the resistance to completion of metadata from?

→ Mistakes or failures are a great way to learn!

Electronic Lab Notebooks

Might be useful to some of you for documentation during the project. Check with your local institution to see what is offered. There are also databases

https://datamanagement.hms.harvard.edu/analyze/electronic-lab-notebooks

https://www.data.cam.ac.uk/data-management-guide/electronic-research-notebooks/electronic-research-notebook-products

Features	Specifications										
		Arxspan	Benchling	BIOVIA	Chemotion	Confluence	Docollab	ecLabNote	eLabFTW	eLabJournal	ELOG	Evernote
Interactivity											
Search functions can search across file formats and beyond typos		✓	*	*		*	*	No response received	*	*	*	*
Ability to manipulate files and images		✓	*	No response received	*	*	*	No response received	✓	✓	*	No response received
Support for multiple open windows		✓	✓	*	✓	✓	✓	No response received	✓	✓	✓	✓
Ability to link out		✓	✓	No response received	✓	*	✓	✓	✓	✓	✓	✓
Support for Researcher Documentation											
Hyperlink support		✓	✓	No response received	✓	✓	✓	✓	✓	✓	✓	✓
Metadata Creation Prompts		✗	✓	No response received	✓	✗	✓	No response received	✗	✗	✓	✗
Rights Management (licensing)		✓	✗	No response received	*	*	✓	No response received	✗	✓	*	No response received
Protocol Integration		✓	✓	✓	✓	✓	✓	No response received	✓	✓	*	*
Adaptability to Lab workflows											
Accounts/Permissions Levels		✓	✓	No response received	✓	*	✓	✓	✓	✓	✓	✓
Internal Data Sharing		✓	✓	*	✓	✓	✓	No response received	✓	✓	✓	✓
Adaptable to a Variety of Workflows		*	*	No response received	No response received	*	*	No response received	✓	No response received	No response received	No response received
Compatibility with authoring tools		*	✓	No response received	*	*	*	No response received	✗	✓	✗	No response received
Windows Compatible		✓	✓	No response received	✓	✓	✓	✓	✓	✓	✓	✓
Macintosh Compatible		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Linux Compatible		✓	✓	✗	✓	✓	✓	No response received	✓	✓	✓	No response received
Android Compatible		✓	✓	✓	✓	✓	✓	No response received	✓	✓	✓	✓
iOS Compatible		✓	✓	✓	✓	✓	✓	No response received	✓	✓	✓	✓

						commercial bio/pharma customers.
Benchling	I, G	B	VC, D, G	F		Free (with capacity limitations) for academic users, user-friendly, self-contained cloud service, Molecular Biology bias. CRISPR tools.
Biovia	G, D	W, M	VC, LS	X		ELN product has basic but robust feature set and workflow, strong in compliance, deployed campus-wide at some institutions. Part of a large suite of BIOVIA products.
BrightLab	I, G, D	B, I, A	VC	F, X		Attractive interface, versatile, available for academia and industrial labs. Instrument integration and LIMS functionality, including direct Sigma-Aldrich ordering.
CDD Vault (Collaborative Drug Discovery)	G, D	B	Not specified	X		ELN module is integrated into LIMS interface and workflow leans towards chemistry and very data-driven documentation requirements. Good, flexible collaboration features, and good integration with MS Office.
c6h6 (Cheminfo ELN)	I, G, D	B (Chrome)	VC, PC, LS	F, O		Free, community-developed tool focused on chemistry with support for many characterisation techniques. FAIR data at the core, i.e. all data are converted into a standard form upon import and can be exported to Zenodo. Source code available. (NB supervisor feature needs suitable groups to be set, for which there is currently no convenient frontend.)

Template

ProcessDataJournal.txt - Notepad

File Edit Format View Help

Processing Journal for Experiment XX

---Original Files---
Created by:
Date(s) Created:

Restrictions:

Notes:

---Processing Updates----
Files(s):
Edited by:
Date:
Software Used:
What was done:

Example

Processing Journal for Experiment XX Data Cleanup

---Original Files---
Files: ornaments_folder
Created by: ▉▉▉▉
Date Created: 10.11.2016

Restrictions: focus on the left hand fingers!

Notes: I will rename all the files starting with the number or the take and then what
       each file focusses on. I will remove the special caracters off the name.
       I renamed a few markers (head) to be current with the plug in gait marker set
       (as much as posible). See the cello_marker_set.xlsx file.

---Processing Updates----

File: 0011_accent_mordant.qtm
Edited by: ▉▉▉▉
Date: 18.12.2020
Software Used: Qualsis

What was done: Markers are labelled throughout the file. All the bones are also created.
       Gap filled the finger of the left hand up to 20 frames, and everything else
       up to 40 frame. I also tried up to 50 and 60 for the less important markers, but
       there is no improvement.

Quality of data, notes: It has a decent filled trajectory for the fingers of the left
       hand. Might be worth analysing further.

RITMO Processing Journal - Work in Progress

Summary

Getting started with metadata can be overwhelming. There is no perfect solution and the best one is often individual.
You're not alone! There are a lot of resources out there that can help.
Here is a systematic approach, I probably missed some stuff but hope it helps anyway
Also a ton of links

More Reading:

DataOne:
https://dataoneorg.github.io/Education/lessons/07_metadata/
07_metadata.pdf
CESSDA:
https://www.cessda.eu/Training/Training-
Resources/Library/Data-Management-Expert-Guide/2.-
Organise-Document/Documentation-and-metadata
RDM Kit from Elixer:
https://rdmkit.elixir-europe.org/metadata_management.html
FAIR Data Collective Generic Metadata Set Template:
https://github.com/fair-data-collective/generic-dataset-
metadata-template
RITMO:
https://www.uio.no/ritmo/english/
https://www.uio.no/ritmo/english/research/labs/fourms/handbo
ok/data-management/