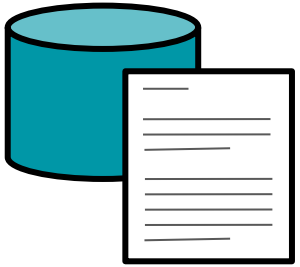
SCIENCE MEETS LIFE

Documentation and Metadata

Alexander Botzki and Bruna Piereck

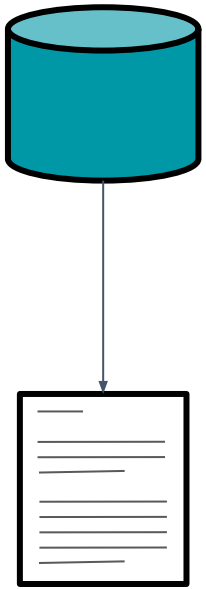
VIB

Which kinds of data/documents we have?



Samples
Protocols / Resources of Products
Images
Code
Data (e.g. readouts - various file formats, …)
Presentations, Posters
Articles
Documentation

Data vs Metadata

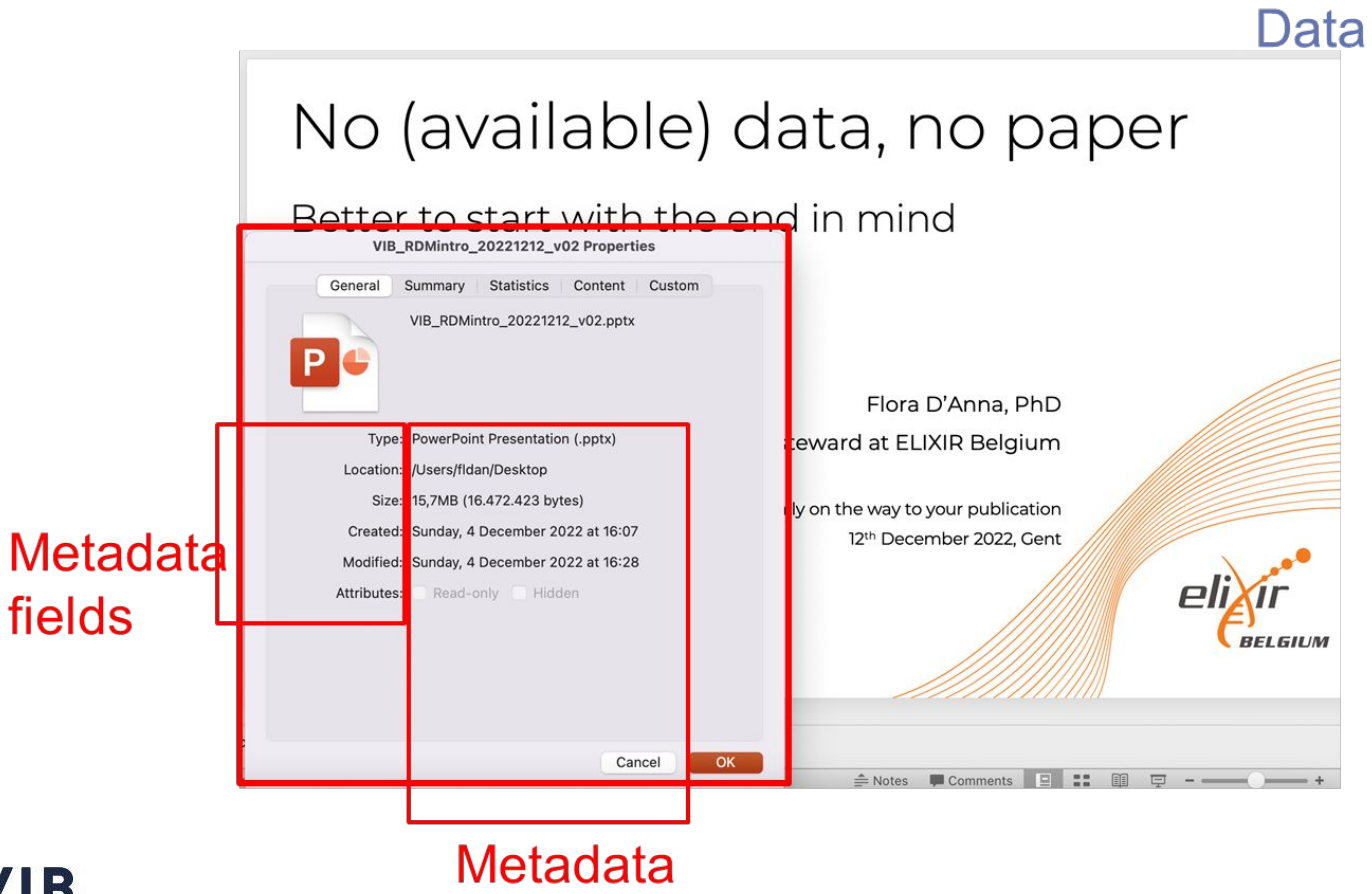


Data itself

Data descriptor

(metadata)

Example borrowed from the session "No (available) data, no paper" by Flora D'Anna



SCIENCE MEETS LIFE

General consideration with the publishing in mind

1. Write the documentation in such a way that someone else who is known to the field can **not mis-interpret** any of the data, even if they tried.

General consideration with the publishing in mind

2. Documentation at two levels

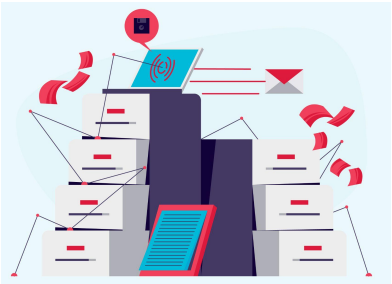
- **Project/Study level:**

Title	Summary
Aims	Authors
Funds	Methods
License	Data sets ID
Folder structure	
File naming	

By pch.vector on freepik.com

- **Data-level:**

documentation



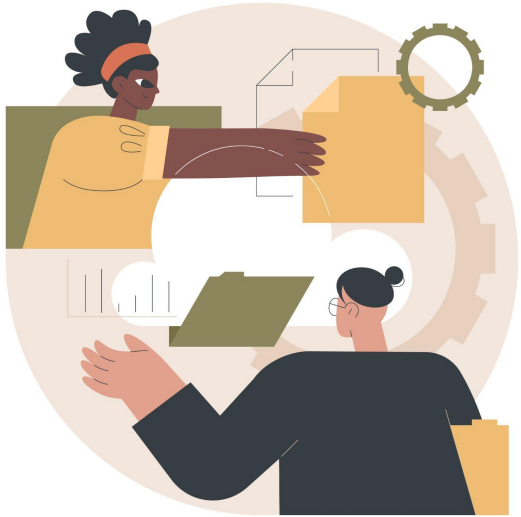
By freepik

https://rdmkit.elixir-europe.org/metadata_management

SCIENCE MEETS LIFE

General consideration with the publishing in mind

3. Both the *study*- and *data*-level documentation must be:
   - Generated **as early as possible** in the research process
   - **Maintained**, in order to be accurate and complete



By vectorjuice on freepik.com

Why would I use an ELN for documentation?

Paper replacement
Project-based system

Share data
Experiments, protocols, …

Standardization of experiments
Templates

Intellectual property protection
Data stored in central database

Eventually Office integration
MS Office: Word, Excel, PowerPoint

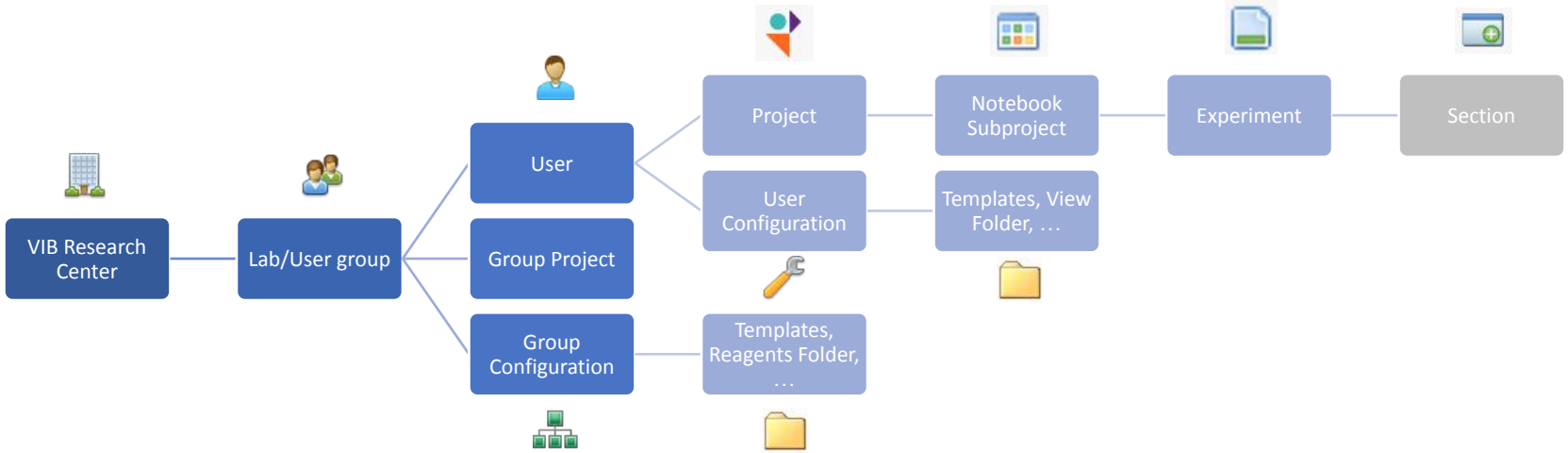
Search function
Advanced search

eLab journal - 19Nov - Leuven
eLab journal - 09Dec - Gent

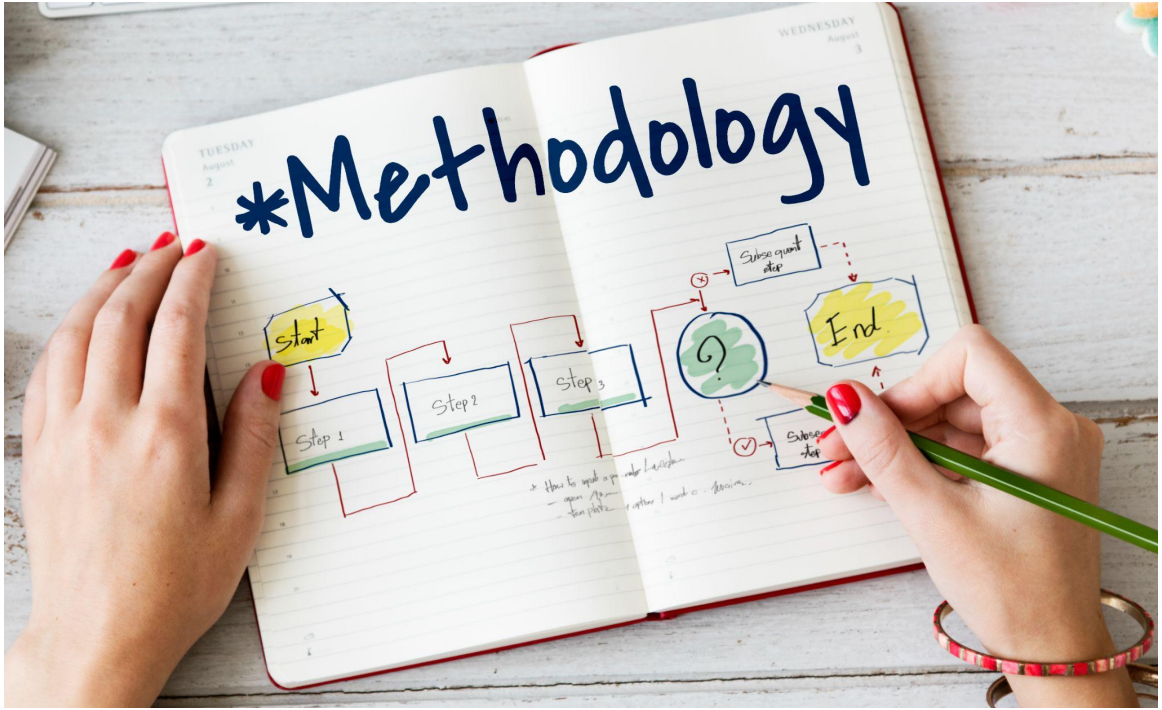
VIB

SCIENCE MEETS LIFE

Project-based system



Looking into one document: My material and methods



By rawpixel.com at frepik.com

How to write useful documentation?

- Exercise:
  - In groups
    - Wet Lab
    - Data Analysis
- You will analyse a document and define what could be improved to actually achieve reproducibility.

How to write useful documentation?

- Exercise: wet lab protocols

The resulting emulsion was collected in aliquots of 50 µL total volume and thermocycled according to the RT program (42°C for 90 min, 11 cycles of [50°C for 2 min, 42°C for 2 min], 85°C for 5 min, followed by a final hold on 4°C). 125 µL of recovery agent (20% PFO in HFE), 55 µL of GITC Buffer (5 M GITC, 25 mM EDTA, 50 mM Tris-HCl pH 7.4) and 5 µL of 1 M DTT was added to each separate aliquot of 50 µL thermocycled emulsion and incubated on ice for 5 min.

- Exercise: data analysis scripts

Barcode reads were trimmed to exclude the intersub-barcode linear amplification adapters using a mawk script. Reads were then mapped and cell-demultiplexed using STARsolo (Kaminow et al., 2021) in CB_UMI_Complex mode. The resulting STARsolo-filtered count matrices were further analyzed using Scanpy (Wolf et al., 2018). In short, cells were filtered on expression of a maximum of 4000 genes, and a maximum of 1% UMIs from mitochondrial genes. Genes were filtered on expression in a minimum of three cells. Potential cell doublets were filtered out using a Scrublet (Wolock et al., 2019) threshold of 0.25.

Discussion of results:

Annex exercise 2

- https://www.protocols.io/view/hydrop-rna-v1-0-dm6gpwqjjlzp/v2?step=5

  While RT is on, you can prepare GITC buffer and other components necessary in the later steps.

  Droplet Breaking and Purification

- Now, we will break all the droplets and purify the cDNA, which is the result of our in-droplet reverse transcription reaction.
- https://www.protocols.io/view/hydrop-rna-v1-0-dm6gpwqjjlzp/v2?step=6.1

How to document scripts and code

- Exercise 2 part 2 data analysis scripts

Barcode reads were trimmed to exclude the intersub-barcode linear amplification adapters using a mawk script. Reads were then mapped and cell-demultiplexed using STARsolo (Kaminow et al., 2021) in CB_UMI_Complex mode. The resulting STARsolo-filtered count matrices were further analyzed using Scanpy (Wolf et al., 2018). In short, cells were filtered on expression of a maximum of 4000 genes, and a maximum of 1% UMIs from mitochondrial genes. Genes were filtered on expression in a minimum of three cells. Potential cell doublets were filtered out using a Scrublet (Wolock et al., 2019) threshold of 0.25.

Data analysis scripts under version control

The best possible way of documenting scripts and code for software is using git and github (or alternative online 'backup' solution to github).

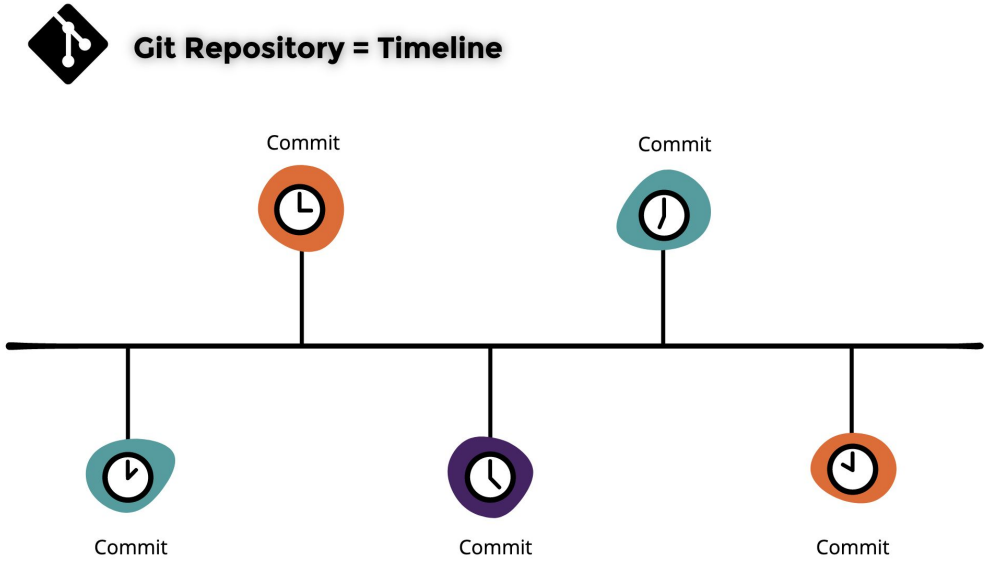


Illustration borrowed from the "Introduction to Git & GitHub" course by Bruna Pierecka and James Collier, ccby 4.0

SCIENCE MEETS LIFE

Data analysis scripts under version control

The best possible way of documenting scripts and code for software is using git and github (or alternative online backup solution to github).

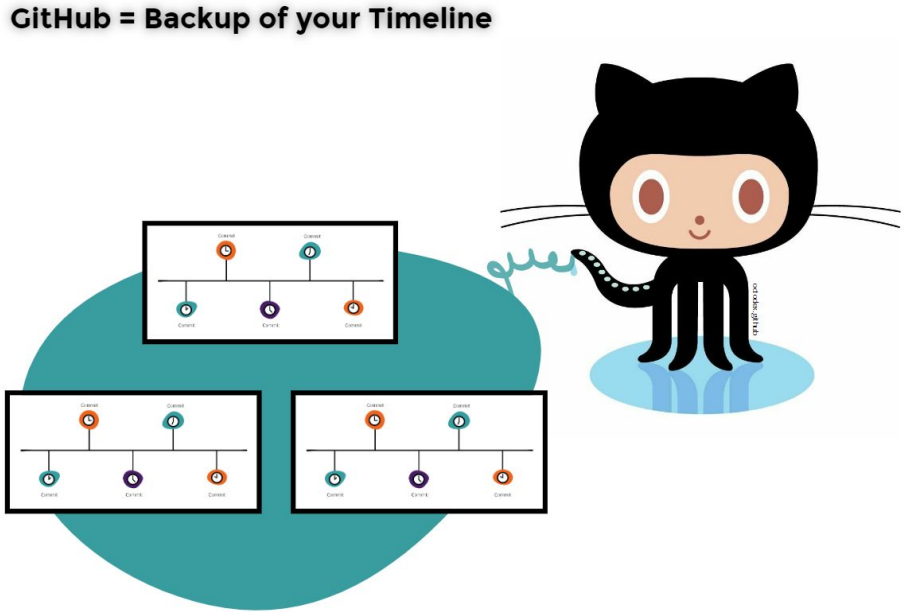


Illustration borrowed from the "Introduction to Git & GitHub" course by Bruna Pierecka and James Collier, ccby 4.0

SCIENCE MEETS LIFE

Data dictionaries and codebooks

Let's explore this example

The association between rainforest disturbance and recovery, tree community composition, and community traits in the Yangambi area in the Democratic Republic of the Congo

Information and data analyses and plots:
https://zenodo.org/record/6979778#.Y5d1ZbLMKeM
Publication:
https://doi.org/10.1017/S0266467422000347

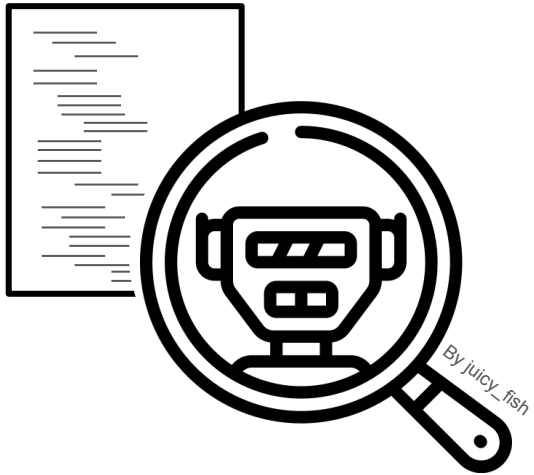
What's in the paper? Zoom in on metadata

- **Data availability**
  The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE175684
- A copy is available in the European Nucleotide Archive at PRJNA733185

What do you know about metadata?

Metadata

Human findable

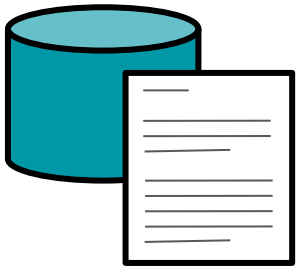


Machine findable

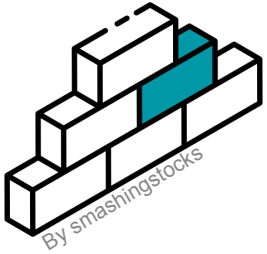
Slide borrowed from "Metadata and Findabilty" session, by Bruna Piereck and Alexander Botzki; from "FAIR material made by design" course, ccbysa

SCIENCE MEETS LIFE

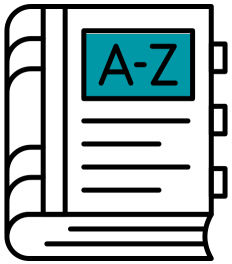
What is Metadata



Data to describe Data



Structured data



By Iconsea

Descriptor with clear meaning

By Freepik

“Love note to the future about data”

How is Metadata structured

Machine findable

Type

Profile

Recipe

A Schema.org Type

Thing > CreativeWork > HowTo > Recipe

[more...]

A recipe. For dietary restrictions covered by the recipe, a few common restrictions are enumerated via suitableForDiet. The keywords property can also be used to add more detail.

Property	Expected Type	Description
Properties from Recipe		
cookTime	Duration	The time it takes to actually cook the dish, in ISO 8601 duration format.
cookingMethod	Text	The method of cooking, such as Frying, Steaming, ...
nutrition	NutritionInformation	Nutrition information about the recipe or menu item.
recipeCategory	Text	The category of the recipe—for example, appetizer, entree, etc.
recipeCuisine	Text	The cuisine of the recipe (for example, French or Ethiopian).
recipeIngredient	Text	A single ingredient used in the recipe, e.g. sugar, flour or garlic. Supersedes ingredients.
recipeInstructions	CreativeWork or ItemList or Text	A step in making the recipe, in the form of a single item (document, video, etc.) or an ordered list with HowToStep and/or HowToSection items.
recipeYield	QuantitativeValue or Text	The quantity produced by the recipe (for example, number of people served, number of servings, etc).
suitableForDiet	RestrictedDiet	Indicates a dietary restriction or guideline for which this recipe or menu item is suitable, e.g. diabetic, halal etc.
Properties from HowTo		
estimatedCost	MonetaryAmount or Text	The estimated cost of the supply or supplies consumed when performing instructions.
performTime	Duration	The length of time it takes to perform instructions or a direction (not including time to prepare the supplies), in ISO 8601 duration format.
prepTime	Duration	The length of time it takes to prepare the items to be used in instructions or a direction, in ISO 8601 duration format.
	CreativeWork or HowToSection or	A single step item (as HowToStep, text, document, video, etc.) or a HowToSection. Supersedes steps

Type

Profile

Recipe

A Schema.org Type

Thing > CreativeWork > HowTo > Recipe

[more...]

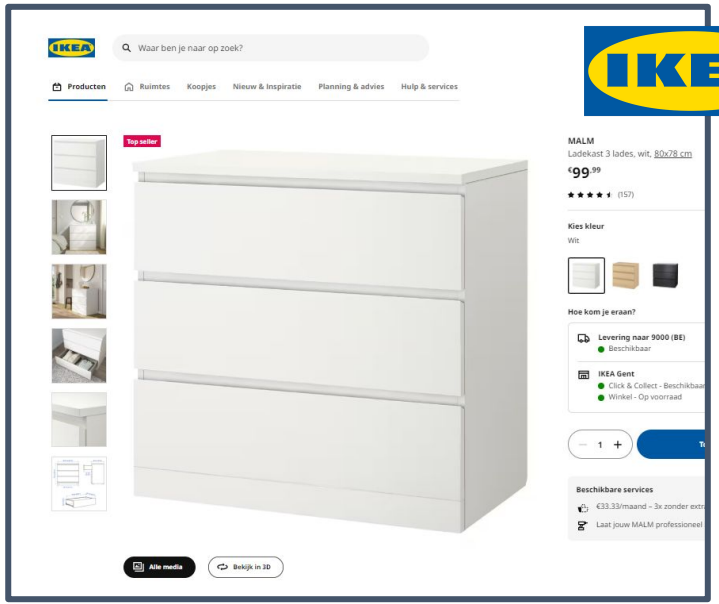
A recipe. For dietary restrictions covered by the recipe, a few common restrictions are enumerated via suitableForDiet. The keywords property can also be used to add more detail.

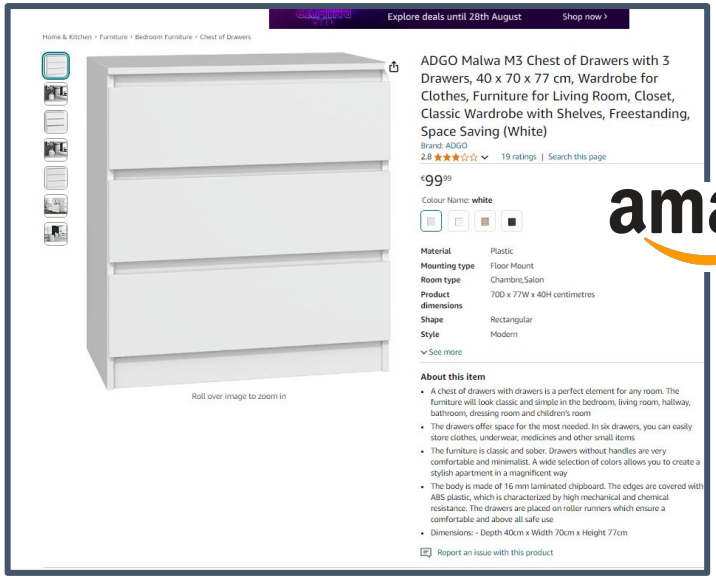
Property	Expected Type	Description
Properties from Recipe		
cookTime	Duration	The time it takes to actually cook the dish, in ISO 8601 duration format.
cookingMethod	Text	The method of cooking, such as Frying, Steaming, ...
nutrition	NutritionInformation	Nutrition information about the recipe or menu item.
recipeCategory	Text	The category of the recipe—for example, appetizer, entree, etc.
recipeCuisine	Text	The cuisine of the recipe (for example, French or Ethiopian).
recipeIngredient	Text	A single ingredient used in the recipe, e.g. sugar, flour or garlic. Supersedes ingredients.
recipeInstructions	CreativeWork or ItemList or Text	A step in making the recipe, in the form of a single item (document, video, etc.) or an ordered list with HowToStep and/or HowToSection items.
recipeYield	QuantitativeValue or Text	The quantity produced by the recipe (for example, number of people served, number of servings, etc).
suitableForDiet	RestrictedDiet	Indicates a dietary restriction or guideline for which this recipe or menu item is suitable, e.g. diabetic, halal etc.
Properties from HowTo		
estimatedCost	MonetaryAmount or Text	The estimated cost of the supply or supplies consumed when performing instructions.
performTime	Duration	The length of time it takes to perform instructions or a direction (not including time to prepare the supplies), in ISO 8601 duration format.
prepTime	Duration	The length of time it takes to prepare the items to be used in instructions or a direction, in ISO 8601 duration format.
	CreativeWork or HowToSection or	A single step item (as HowToStep, text, document, video, etc.) or a HowToSection. Supersedes steps

How does metadata works?

White chest of drawer with **3 drawers**,
depth between 40-50 cm,
width between 70-80 cm,
height between 75-80 cm.

How does metadata works?





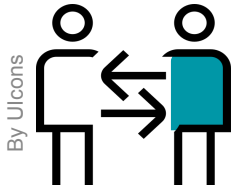
amazon

VIB

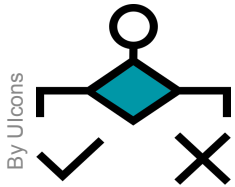
SCIENCE MEETS LIFE

Metadata impact

Findability: Findable by search engines



Sharing: Structured information to evaluate



Decision making: Targeted judgment
Evaluation on reusability

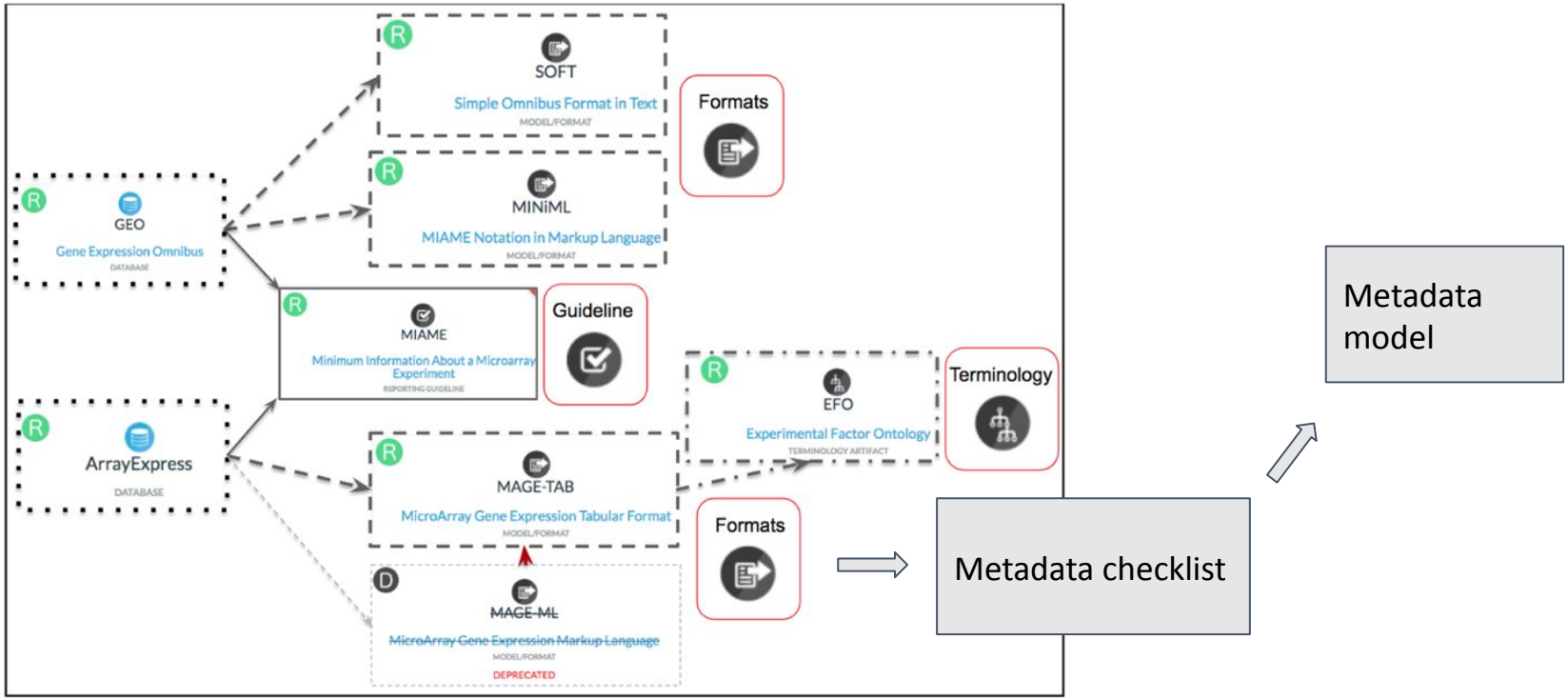
Ikea Ontologies

Let’s watch how Ikea approaches metadata.

From metadata to data models

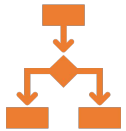
Let’s watch how Ikea approaches metadata.

First a short excursion to nomenclature



https://doi.org/10.1038/s41597-022-01707-6

There are many metadata models



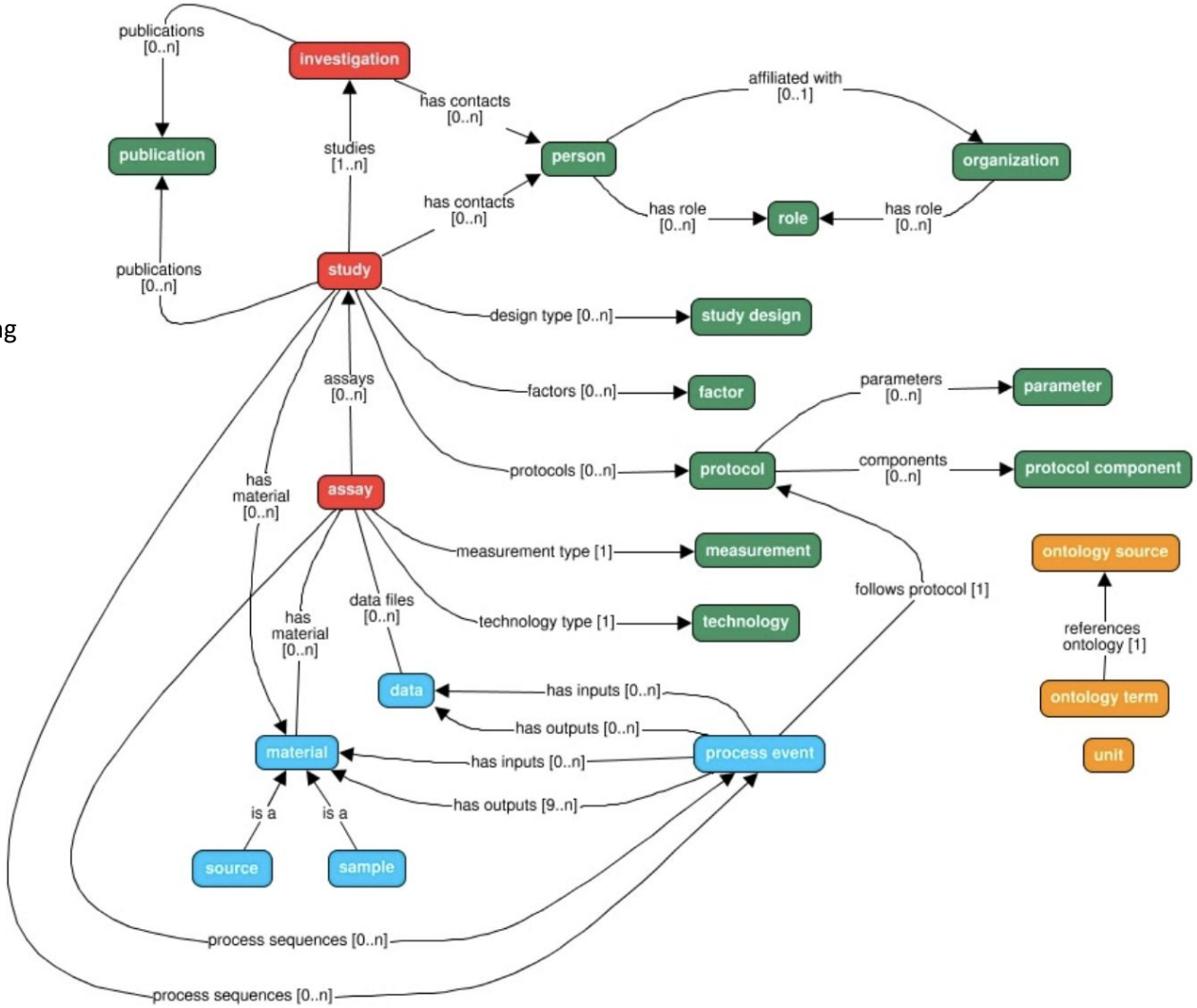
A **metadata model** is a structured framework that defines the types, relationships, and attributes of metadata within a specific context.

Essentially, it is a model that describes how metadata is organized and managed, providing a blueprint for how data about data is structured and utilized.

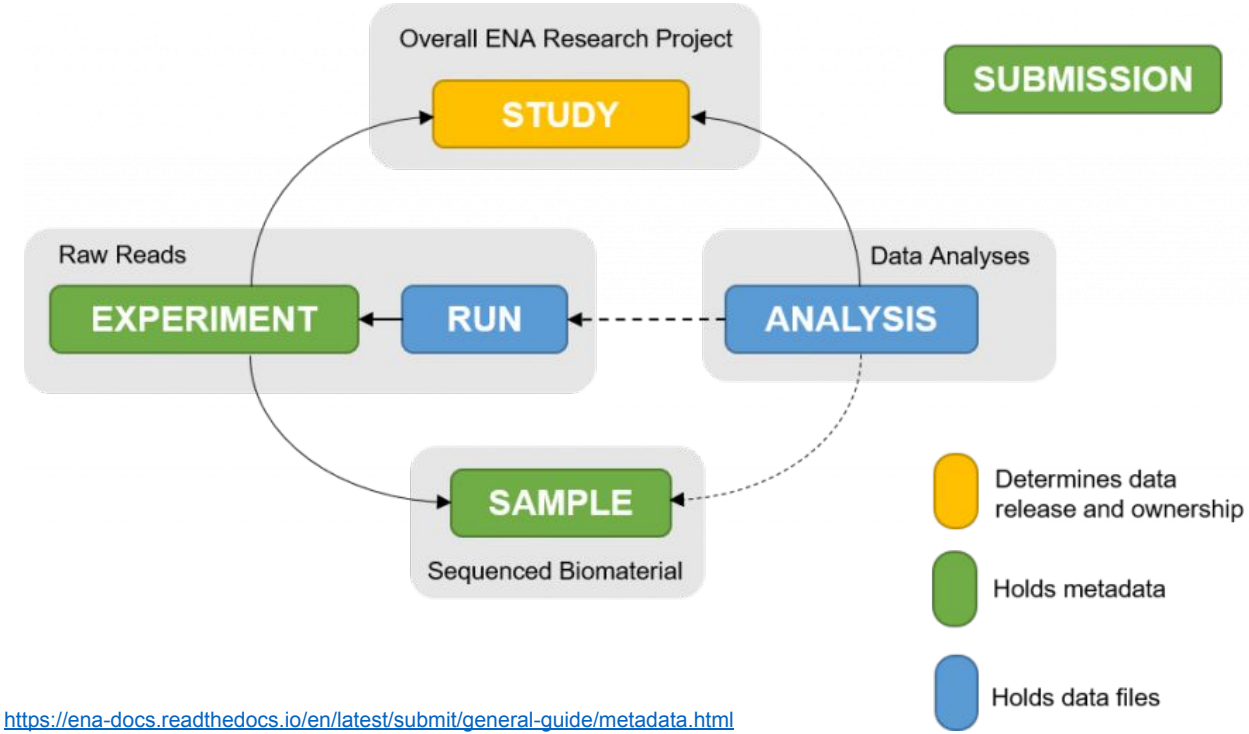
E.g. https://faircookbook.elixir-europe.org/content/recipes/infrastructure/dats-model.html

There are many metadata models

https://isa-tools.org/isa-api/_images/isa-model.png

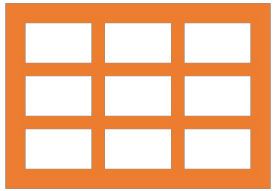


ENA's metadata model



https://ena-docs.readthedocs.io/en/latest/submit/general-guide/metadata.html

Let’s focus on the metadata checklists



- From ENA - their checklist of metadata
- ELIXIR's Fair Cookbook for transcriptomics data

- Exercise 3 Let’s fill in the metadata based on the paper

External references

- Harvard Comparison Grid: https://datamanagement.hms.harvard.edu/electronic-lab-notebooks
- NKI: LabGuru for Wetlab and Castor for Clinical Data Capature
- MDC and BIH – LIMS is available, RDM SODAR (Omics data)

Documentation and support

VIB & ELIXIR Support

VIB Data Core - for e-lab journal (Follow training and Learn VIB support points)
- eLab journal - 19Nov - Leuven
- eLab journal - 09Dec - Gent

RDM more specific support should exist per center, taking their particularities in account

ELIXIR services - usegalaxy.eu

More information at: https://www.elixir-belgium.org/services

Data Core Team Lead Data Management: flora.danna@vib.be

What did we cover today.

- implement SOP type of approach for your daily documentation of experiments
- describe the impact of documentation on the publication preparation
- make versioning more persistent by using protocols.io and/or your Electronic Lab Notebook
- use github for scripts and code
- apply at least minimal metadata standards for domain-specific data

Our guide for today's session

HyDrop-RNA single-cell library preparation

- https://pubmed.ncbi.nlm.nih.gov/35195064/

- Exercise 1 Documentation at project level