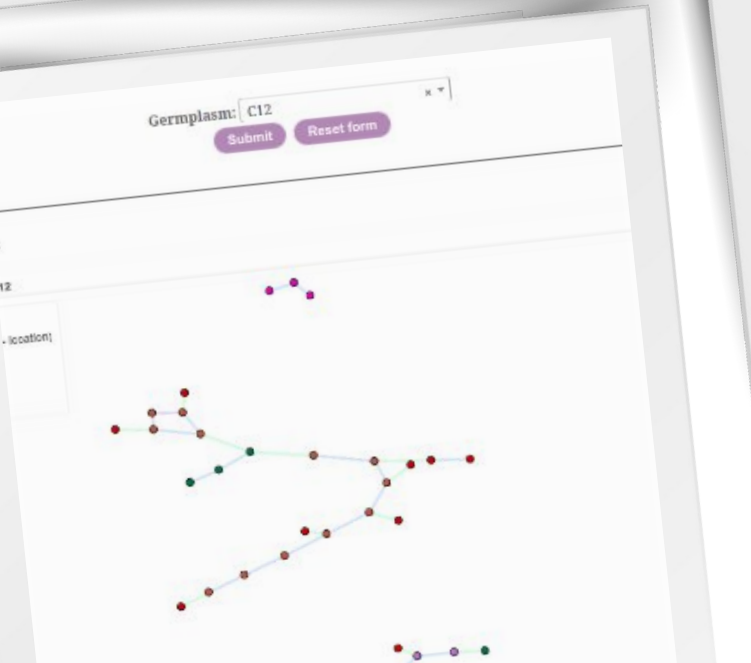


Training in organic breeding

Module 2: Phenomics: approaches and tools for genetic resources and breeding material characterization

Unit 2.3: Guidelines and examples of good practices in data management

Authors: Yannick de Oliveira, Isabelle Goldringer

CC BY SA

Co-funded by the European Union

Co-funded by the European Union

Funded by the European Union, the Swiss State Secretariat for Education, Research and Innovation (SERI) and UK Research and Innovation (UKRI).

UK Research and Innovation

Training in organic breeding organized in 5 Modules

1. **Module 1** - Plant Genetic Resources (PGRs): collection, conservation and exchange to support the increase of agrobiodiversity in farming systems
2. **Module 2 - Phenomics: approaches and tools for genetic resources and breeding material characterisation - FEBRUARY 3rd 2025, 9:00 to 17:30 CET**
3. **Module 3** - Breeding methods fundamentals - FEBRUARY 13th 2025, 9:00 to 18:00 CET
4. **Module 4** - Development and application of molecular methods in organic breeding - MARCH 4th 2025, 9:00 to 18:00 CET
5. **Module 5** - Organic heterogeneous material (OHM) design and development - MARCH 7th 2025, 9:00 to 18:00 CET

LiveSeeding

February 3rd 2025 - 9:00 to 17:30 CET

Unit 2.1: Main descriptors used worldwide in characterizing plant genetic resources

- 9:00-10:30 - UPV (Adrian Rodríguez-Burruezo)
- 10:30-11:00 Break

Unit 2.2: Intro to ShineMas: a web tool dedicated to Seed Lots History, Phenotyping and Cultural Practices[1]

- 11:00-12:30 - INRAe (Yannick de Oliveira, Isabelle Goldringer)
- 12:30-14:00 Lunch Break

Unit 2.3: Guidelines and examples of good practices in data management

- 14:00-15:30 - INRAe (Yannick de Oliveira, Isabelle Goldringer)
- 15:30-16:00 Break

Unit 2.4: Methods for phenotyping and selection of agronomic traits of interest in organic farming

- 16:00-17:30 - IPC (Pedro Mendes Moreira)

1 - An extra practical session to use the tool with own data is scheduled for FEB 10th (9-12h)

T1.4 Training in Organic Breeding

MODULE 2 – Phenomics: approaches and tools for genetic resources and breeding material characterisation

Unit 2.3: Data management	INRAE

Unit 3 – Guidelines and examples of good practices in data management

Yannick De Oliveira

&

Isabelle Goldringer

INRAE

LiveSeeding

Co-funded by the European Union

Funded by the European Union, the Swiss State Secretariat for Education, Research and Innovation (SERI) and UK Research and Innovation (UKRI). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or REA, nor SERI or UKRI.

UK Research and Innovation

Training outline

- Context of reproducibility crisis in science (10 minutes)
- Data management plan and legal constraint regarding data (15 minutes)
- What is FAIR data (10 minutes)
- Guidelines to manage data (30 minutes) :
  - Tidy data
  - Standards and metadata
  - Vocabulary
  - Licenses
  - Data warehouse
- Short quiz (10 minutes)

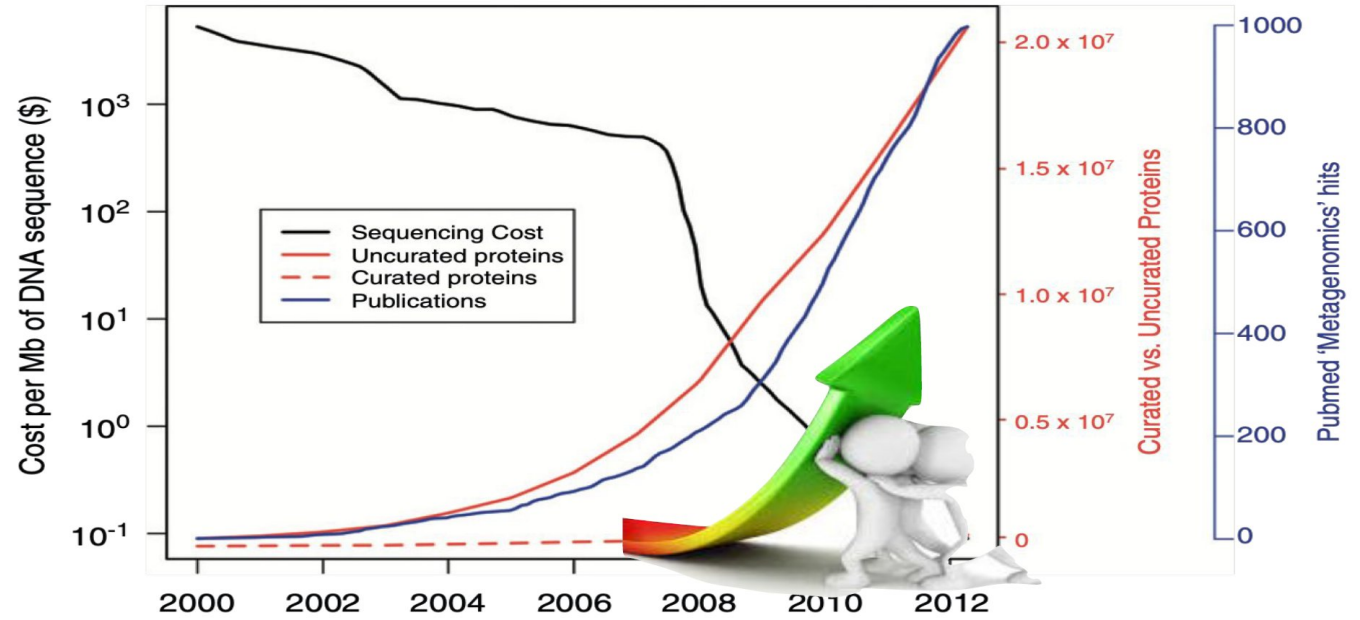
LiveSeeding

Why manage data properly ?

Context

Digital transformation is changing the landscape involving a data deluge and paradigm shift that need to be to manage.

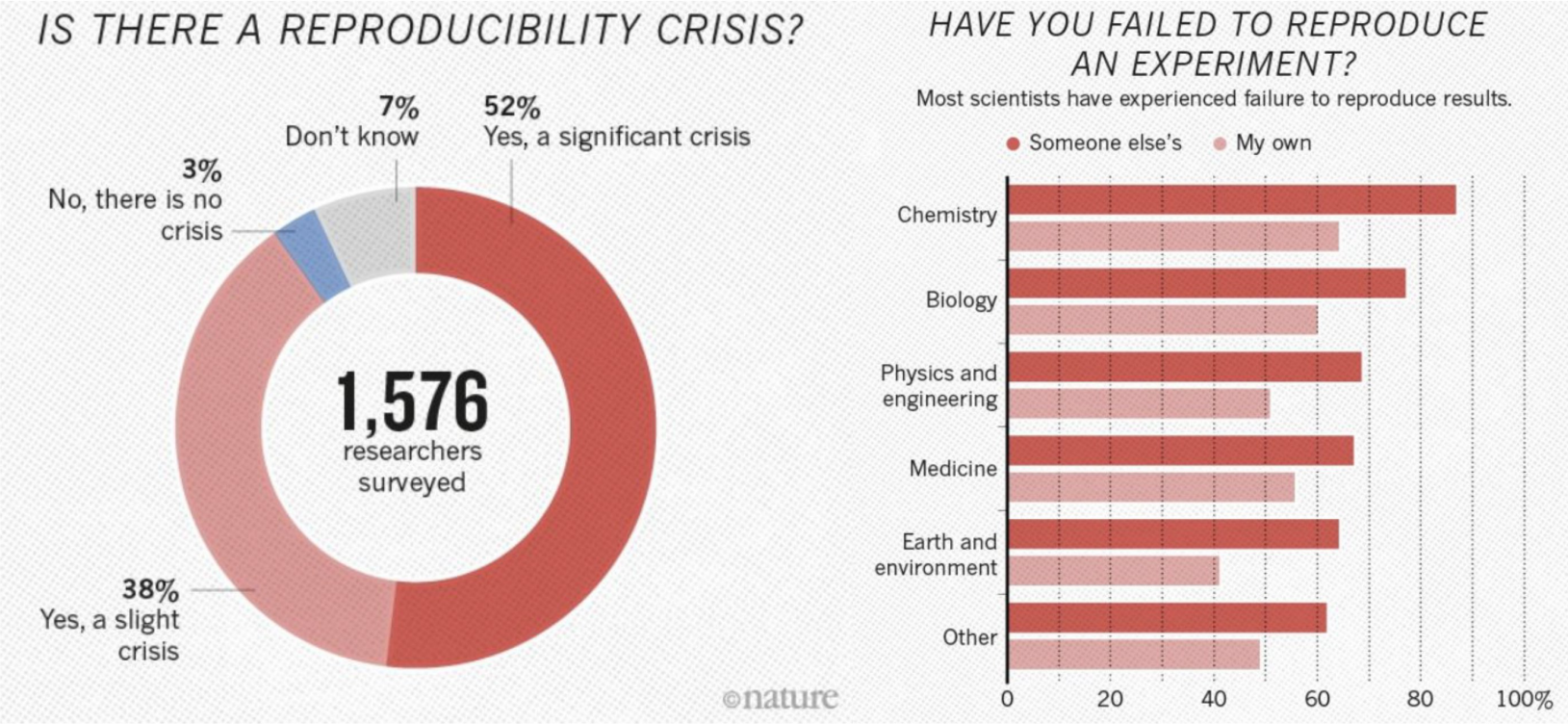
- **Before: experience design > data collected > analysis**
- **Now: data production > organization > analysis > sharing information**



LiveSeeding

Science isn’t reproducible

90% of questioned scientists thinks reproducibility crisis exists



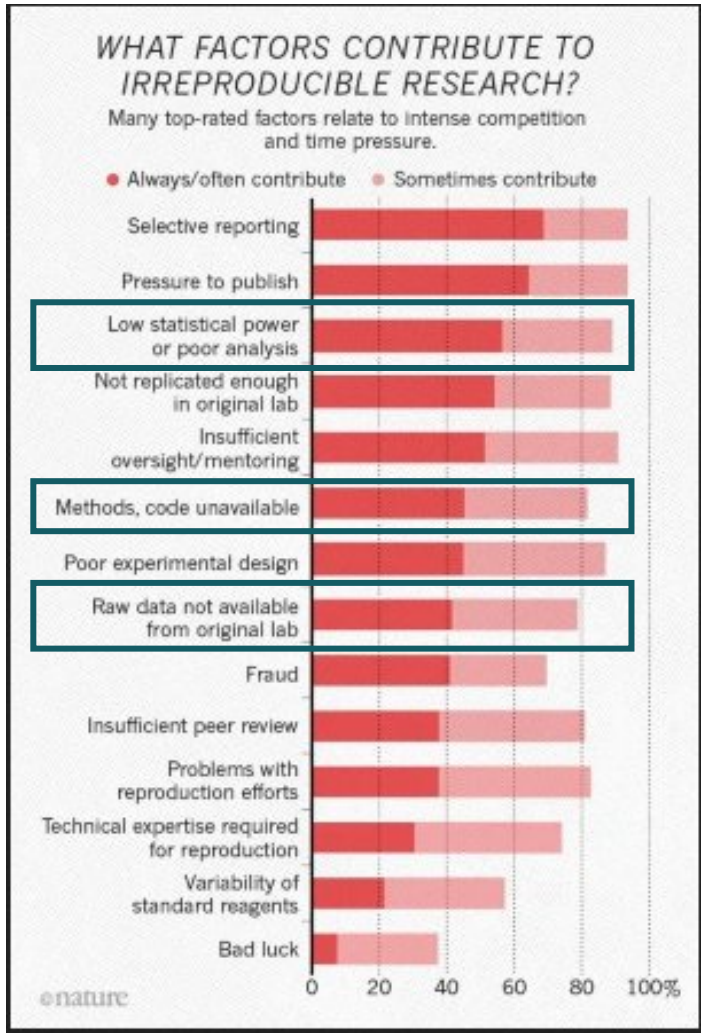
A significant part of them admit they already failed to reproduce their own experiment

"1,500 scientists lift the lid on reproducibility". Nature. 533: 452–454 - 2016

https://doi.org/10.1038/533452a

LiveSeeding

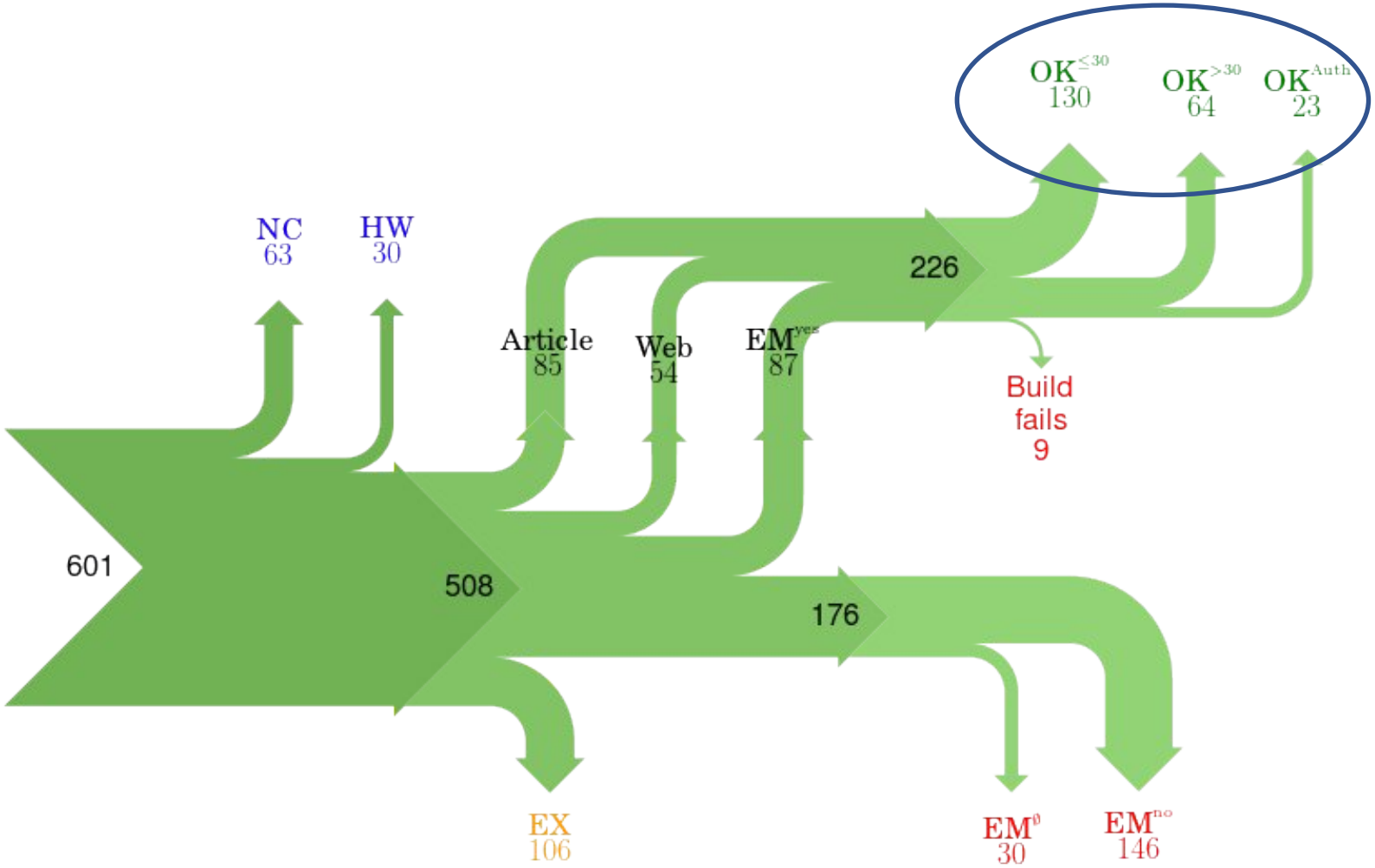
A multifactor crisis



Reasons advanced by scientists are multiple but, a significant part of them are related to data management, their analysis and the tools (source code) used to do these analyses.

LiveSeeding

An example with source code



In this study, only 54% of source code can be build (repeatability ok).

An important part needed significant efforts for building or the help of the author.

In most cases programs can't be build because author refuse to share the code or didn't answer to email.

LiveSeeding

Collberg & al. - Repeatability in computer systems research
https://doi.org/10.1145/2812803

Legal obligation

A report of the EU in 2018 estimate the annual cost of producing non-FAIR data : ~10,2bn€

European Law

- **Open Data Directive (16 July 2019)**
- **Free movement of non-personal data in the EU (2018/1807)**
- **GDPR (2018)**

European Commission

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data
European Commission
Directorate-General for Research and Innovation
Directorate A — Policy Development and Coordination
Unit A.2 — Open Data Policy and Science Cloud
Contact Athanasios Karalopoulos
E-mail Athanasios.Karalopoulos@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

https://data.europa.eu/doi/10.2777/02999

LiveSeeding

Focus on GDPR

Le RGPD en 5 points

- Declaration of ALL personal data processing to the institutional DPO (Data Protection Officer) and assurance of compliance
- Information for data subjects +/- consent
- Secure media and data transfers
- Individual rights (access, modification, deletion and portability)
- PIA (Privacy Impact Assessment) for sensitive data

DPO = Data Protection Officer

- Ensures the compliance of personnal data processing (information of individual, individual rights, personnal data processing register)
- He is NOT responsible

LiveSeeding

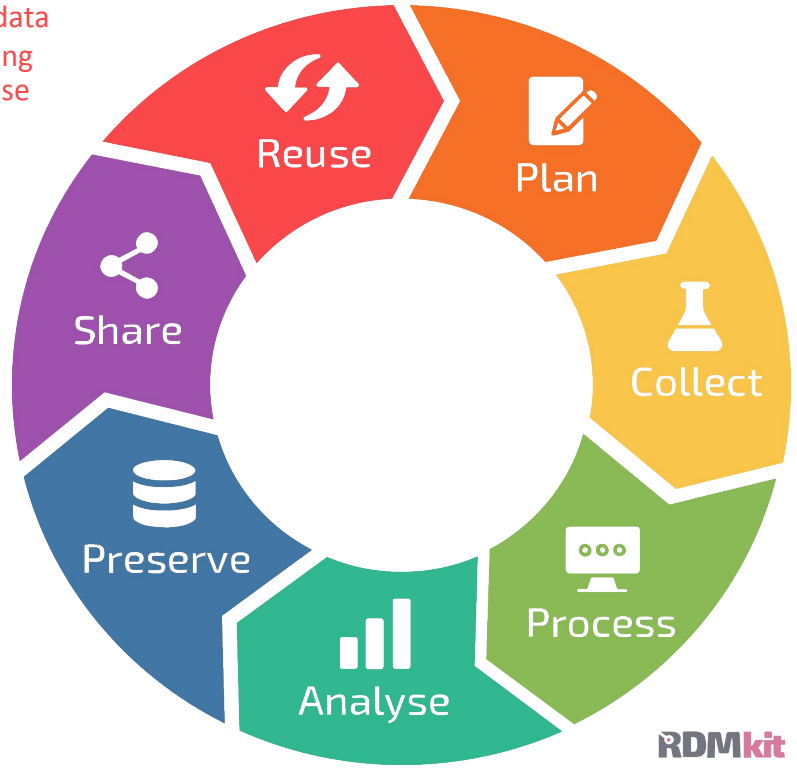
Data Management Plan

Data Management Plan

A DMP is a document that describes how the data of a research project or an entity will be managed throughout its lifecycle

Carry out new research
Re-examine results/data
Teaching and learning
Respect reuse license

Define data access rights
Choose a license



Data collect
Data production

Migrate data to a standard format and suitable media
Complete metadata and documentation
Save and store
Submit data in a data warehouse
Long-term data archiving

Check, Clean, Validate
Store
Exchange data securely

Data analyses
Produce scientific results
Publish articles

LiveSeeding

DMP life cycle

Setting up the project
Section on data management
Evakuate the cost of open science (data manager, storage…)

Starting the project
Setting up the Data Management Plan (DMP)

During the project
Data management following FAIR principle
Update the DMP

End of the project
Share data « ***as open as possible, as closed as necessary*** »
Submit in data warehouse in open access

LiveSeeding

DMP objectives

Implement best practices, respecting FAIR principles (Findable, Accessible, Interoperable, Reusable)

- **Ensure reproducibility of experiments by describing the data and how it was obtained**
- **Enabling data to be understood and reused**
- **Avoid data loss through appropriate storage**
- **Establish roles and responsibilities of everyone**
- **Respect the law and individuals by clarifying the legal and ethical framework**
- **Clarify re-use rights and sharing modality**

LiveSeeding

Project DMP vs Entity DMP

Project DMP

- Project funded
- Specific scope and fixed term
- Mandatory

is based on →

← *help to prepare*

Entity DMP

- Research laboratory, plateform, breeder company etc.
- Larger scope and no fixed term
- Not mandatory

LiveSeeding

Tips : DMP tools

DMP OPIDoR

- Can include template from funders such as Horizon Europe

- Easy to use: web interface with form to fill

—> https://dmp.opidor.fr/

DSW

Data stewardship wizard
https://ds-wizard.org/

LiveSeeding

What is going wrong ?

Click on the link and share you idea on what is going wrong in this video.

https://postit.colibris-outilslibres.org/fairdataliveseeding

https://www.youtube.com/watch?v=66oNv_DJuPc

LiveSeeding

FAIR data

What is it ?

The FAIR Principles are a set of guidelines that aim to make data Findable, Accessible, Interoperable and Reusable.

- **It provides guidelines for scientific data management and are relevant to all stakeholders of the digital ecosystem.**
- **They are aimed directly at data producers and publishers to promote maximum use of scientific data.**
- **They focus on the ability of machines to manage data automatically, with the minimum of human intervention.**

References :

- **Wilkinson 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**
- **Les principes FAIR. DORANum. http://doi.org/10.13143/z7s6-ed26 (french)**

LiveSeeding

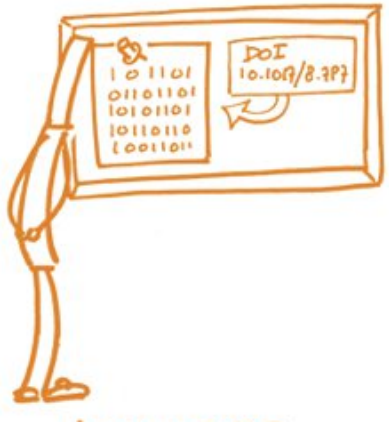
Findable

Facilitate data (and metadata) discovery for both humans and machines

- Data have a PID
- Data are described by metadata
- Thess metadata include the PID of the data they describe
- Data are submited in a data warehouse

LiveSeeding

Accessible



Enable data access and download, which may include authentication and authorization

- **Data can be accessed via a standard communication protocol**
- **The protocol is free and open**
- **This protocol makes possible an access by authentication if required**
- **Metadata remains accessible even if data is not (disappeared or inaccessible)**

LiveSeeding

Interoperable

Enable data exploitation and integration whatever of the IT environment used

- **Data are described using controlled vocabularies**
- **The vocabulary used is complient with FAIR principles**
- **Metadata are contextualized with links to other data**

LiveSeeding

Reusable

Enable data to be reused for future research

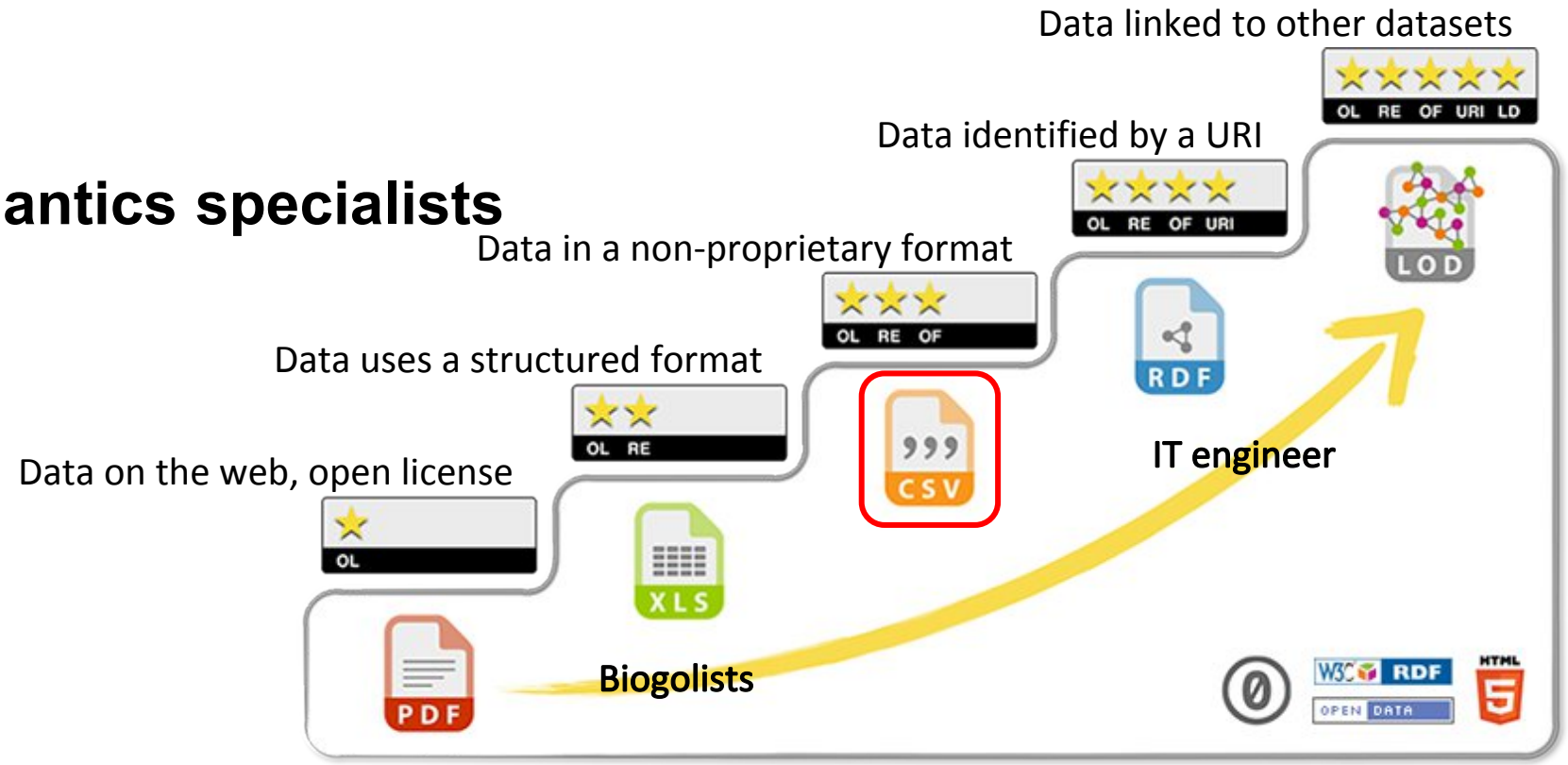
- **Metadata contains all information that may be useful (plurality of attributes)**
- **A license for reuse is assigned to the data**
- **The description of the data indicates its origin**
- **Data sharing follows scientific community standards**

LiveSeeding

The 5 stars of FAIR data

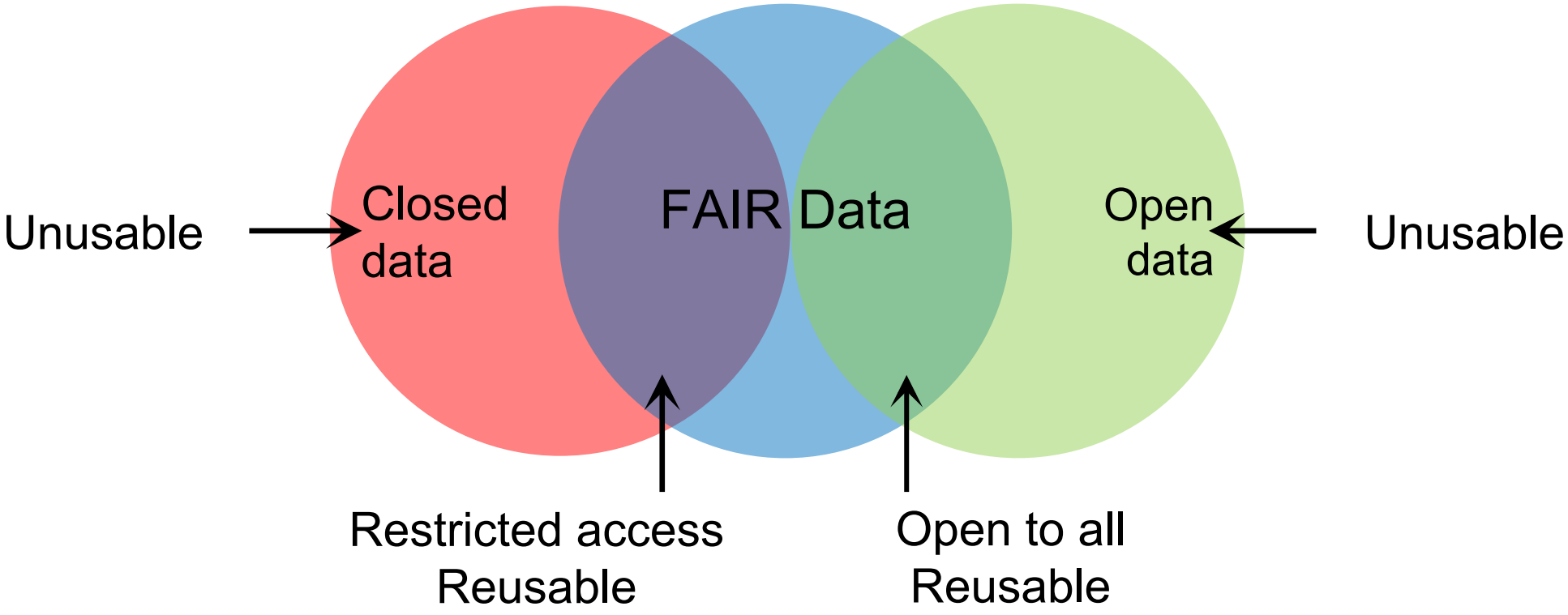
Progress towards FAIR and Open Data requires multidisciplinary cooperation

- **Biologists**
- **IT Engineer**
- **Ontology/semantics specialists**



LiveSeeding

Accessible ≠ Open



LiveSeeding

Guidelines for data management

Tidy data

Tidy data

What is it and what for?

How to move from messy to tidy data?

Tidy data and tidy tools => everything in Whickham (2014)

(The steps further: designing a global dataset)

Additional available material

Tidy data

- 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003)
- Real-world datasets = often few, if any, constraints on their organization, datasets often constructed in bizarre ways
- Data preparation is not just a first step, but must be repeated many times over the course of analysis (new problem arising, new data collected…)
- => Need to structure datasets to facilitate analysis

LiveSeeding

Source : Wickham H (2014)

Tidy data

- Principles : to provide a standard way to organize data values within a dataset.
- => will make initial data cleaning easier, facilitate initial exploration and analysis of the data, and simplify the development of data analysis tools that work well together.
- Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).

LiveSeeding

Source : Wickham H (2014)

Data structure

	Yield_Org	Yield_Non Org
Variety_A	-	40
Variety_B	35	45
Variety_C	27	30

Table 1

	Variety_A	Variety_B	Variety_C
Yield_Org	-	35	27
Yield_Non Org	40	45	30

Table 2

- Table 1 & 2 = same data but different layout

Source : Wickham H (2014)

Data semantics

- Dataset = collection of values (numbers or strings)
- Every value belongs to a variable and to an observation
- A variable contains all values that measure the same underlying attribute across units
- An observation contains all values measured on the same unit across attributes

Source : Wickham H (2014)

Data semantics

- Reorganizing Table 1 to make the values, variables, observations more clear

	Yield_Org	Yield_Non Org
Variety_A	-	40
Variety_B	35	45
Variety_C	27	30

Table 1

	Yield_Org	Prot_Org
Variety_A	-	40
Variety_B	35	45
Variety_C	27	30

Table 3

- Table 1: mix between variables and observations
- Table 3 would be ok

LiveSeeding

Source : Wickham H (2014)

Module 2 – Unit 3 Tidy data: how to move from messy to tidy data?

Tidy data

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table
- ***Messy data*** is any other arrangement of the data.

Table 1

	Yield_Org	Yield_Non Org
Variety_A	-	40
Variety_B	35	45
Variety_C	27	30

Varieties	Farming practice	Yield
Variety_A	Organic	-
Variety_B	Organic	35
Variety_C	Organic	27
Variety_A	Non Organic	40
Variety_B	Non Organic	45
Variety_C	Non Organic	30

Table 4

Table 4 = tidy **Table 1** => each row represents an observation, the result of one Farming practice on one Variety, and each column is a variable.

LiveSeeding

Source : Wickham H (2014)

Tidy data

- 5 most common problems with messy datasets:
  - Column headers are values, not variable names
  - Multiple variables are stored in one column
  - Variables are stored in both rows and columns
  - Multiple types of observational units are stored in the same table
  - A single observational unit is stored in multiple tables

LiveSeeding

Source : Wickham H (2014)

Module 2 – Unit 3 Tidy data: how to move from messy to tidy data?

Example 1 : Column headers are values, not variable names

- This dataset has three variables, *Population*, *Flower colour* and *Frequency* => to tidy Table 5, we need to melt, or stack it = turn columns into rows

Population	Flower_Blue	Flower_purple	Flower_pink	Flower_white
Pop_1	15	25	40	20
Pop_2	0	10	80	10
Pop_3	20	50	25	5

Table 5

Population	Flower_ colour	Frequency
Pop_1	Blue	15
Pop_2	Blue	0
Pop_3	Blue	20
Pop_1	Purple	25
Pop_2	Purple	10
Pop_3	Purple	50
Pop_1	Pink	40
Pop_2	Pink	80
Pop_3	Pink	25
Pop_1	White	20
Pop_2	White	10
Pop_3	White	5

Table 6

LiveSeeding

Source : Wickham H (2014)

Example 2 : Column headers are values not variable names + Multiple variables stored in 1 column

- This dataset has four variables, *Plant*, *Leaf-number*, *Leaf-side*, and *diseased-surface* => to tidy Table 7, we need to melt it + split a single variable *Leaf#-side* into 2 real variables *Leaf-number* and *Leaf-side*

Plant	Lf1-top	Lf1-Bottom	Lf2-top	Lf2-bottom	Lf3-top	Lf3-bottom
$Plant_1$	50	40	30	30	10	5
$Plant_2$	10	10	5	0	0	0
$Plant_3$	25	20	15	10	5	0

Table 7

Source : Wickham H (2014)

Module 2 – Unit 3 Tidy data: how to move from messy to tidy data?

Example: Column headers are values not variable names + Multiple variables stored in 1 column

Plant	Lf1-top	Lf1-Bottom	Lf2-top	Lf2-bottom	Lf3-top	Lf3-bottom
$Plant_1$	50	40	30	30	10	5
$Plant_2$	10	10	5	0	0	0
$Plant_3$	25	20	15	10	5	0

Table 7

Plant	Lf#-side	Diseased surface
$Plant_1$	Lf1-top	50
$Plant_2$	Lf1-top	10
$Plant_3$	Lf1-top	25
$Plant_1$	Lf1-Bottom	40
$Plant_2$	Lf1-Bottom	10
$Plant_3$	Lf1-Bottom	20
$Plant_1$	Lf2-top	30
$Plant_2$	Lf2-top	5
$Plant_3$	Lf2-top	15
$Plant_1$	Lf2-bottom	30
$Plant_2$	Lf2-bottom	0
$Plant_3$	Lf2-bottom	10
$Plant_1$	Lf3-top	10
$Plant_2$	Lf3-top	0
…	…	…

Table 8

Source : Wickham H (2014)

Module 2 – Unit 3 Tidy data: how to move from messy to tidy data?

Example: Column headers are values not variable names + Multiple variables stored in 1 column

- Table 8

Plant	Lf#-side	Diseased surface
$Plant_1$	Lf1-top	50
$Plant_2$	Lf1-top	10
$Plant_3$	Lf1-top	25
$Plant_1$	Lf1-Bottom	40
$Plant_2$	Lf1-Bottom	10
$Plant_3$	Lf1-Bottom	20
$Plant_1$	Lf2-top	30
$Plant_2$	Lf2-top	5
$Plant_3$	Lf2-top	15
$Plant_1$	Lf2-bottom	30
$Plant_2$	Lf2-bottom	0
$Plant_3$	Lf2-bottom	10
$Plant_1$	Lf3-top	10
$Plant_2$	Lf3-top	0

Table 9

Plant	Leaf-nb	Leaf-side	Diseased surface
$Plant_1$	Lf1	top	50
$Plant_2$	Lf1	top	10
$Plant_3$	Lf1	top	25
$Plant_1$	Lf1	bottom	40
$Plant_2$	Lf1	bottom	10
$Plant_3$	Lf1	bottom	20
$Plant_1$	Lf12	top	30
$Plant_2$	Lf12	top	5
$Plant_3$	Lf12	top	15
$Plant_1$	Lf12	bottom	30
$Plant_2$	Lf12	bottom	0
$Plant_3$	Lf12	bottom	10
$Plant_1$	Lf3	top	10
$Plant_2$	Lf3	top	0

Source : Wickham H (2014)

Tidy data

- 5 most common problems with messy datasets:
  - Column headers are values, not variable names
  - Multiple variables are stored in one column
  - Variables are stored in both rows and columns
  - Multiple types of observational units are stored in the same table
  - A single observational unit is stored in multiple tables

=> See Wickham H (2014)

LiveSeeding

Tidy data

One way of organizing variables is by their role in the analysis:

- Fixed variables describe the experimental design (known in advance)
- => should come first
- Measured variables are what we measure in the study
- => should follow (related variables contiguous)

LiveSeeding

Source : Wickham H (2014)

Tidy tools

Take tidy datasets as input and return tidy datasets as output

=> the output of one tool can be used as the input to another

Everything you need to know about tidy data and tidy tools is here:

- Wickham H (2014). Tidy Data. Journal of Statistical Software 59. https://doi.org/10.18637/jss.v059.i10.
- Article, codes and examples: https://www.jstatsoft.org/article/view/v059i10

LiveSeeding

Source : Wickham H (2014)

Module 2 – Unit 3 The steps further: designing a global dataset

Mahmoud, R et al. A workflow for processing global datasets: application to intercropping. Peer Community Journal, Volume 4 (2024), article no. E24.

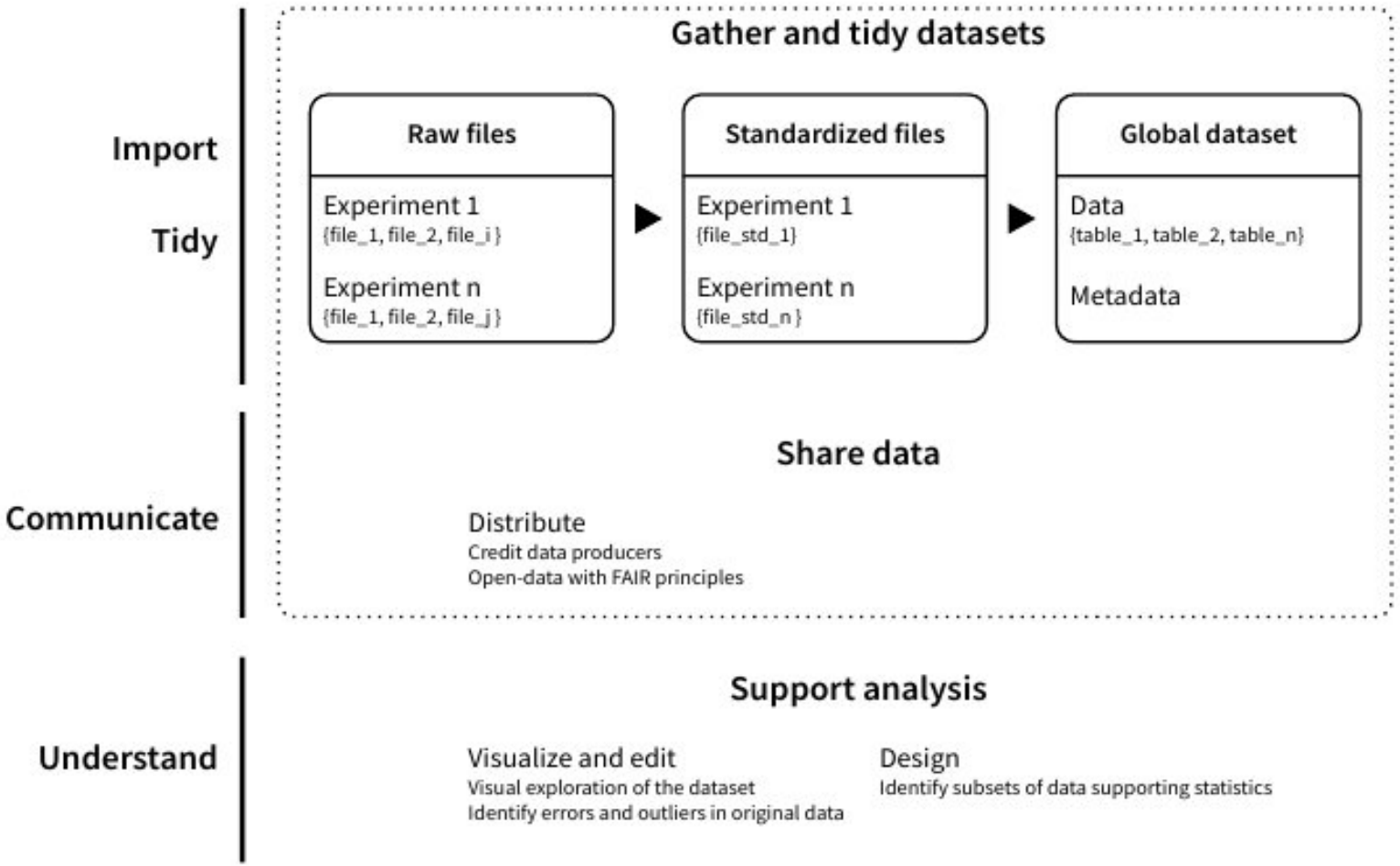


Figure 1 - Main steps for designing global datasets. The left column corresponds to a classical data science workflow. We adapted these steps for global dataset design specificities, to illustrate the importance of data gathering, tidying, and sharing (dotted frame). While some actions supporting subsequent data analysis are generic (visualization, editing), most depend on the chosen analysis strategy.

LiveSeeding

Controlled vocabulary

LiveSeeding

Why controlled vocabularies ?

A controlled vocabulary makes possible to describe concepts used by a community in the same way.

It can be a thesaurus, a glossary, an ontology etc.

A typical (and simple) example of non-controlled vocabulary, a same trait written in different ways : *plant_height*, *plant height*, *plants_height* etc.

LiveSeeding

Thesaurus

Agrovoc (FAO) https://agrovoc.fao.org/browse/agrovoc/en/

Alphabetical | Hierarchy | Groups

- design
- destruction of animals
- detection
- diagnosis
- disinfection
- dispute settlement
- diving
- dredging
- economic activities
  - agriculture
    - agricultural practices
    - farming systems
      - agroforestry
      - agroforestry systems
      - agropastoral systems
      - alternative agriculture
        - biodynamic agriculture
        - **organic agriculture**
        - permaculture
      - aquaculture systems
      - aquatic agricultural systems

systems > farming systems > alternative agriculture > organic agriculture

... > economic activities > agriculture > farming systems > alternative agriculture > organic agriculture

PREFERRED TERM	**organic agriculture**
BROADER CONCEPT	alternative agriculture (en)
RELATED CONCEPTS	organic certification (en)
ENTRY TERMS	*organic farming (en)*
SCOPE NOTE	Agricultural methods without use of chemical products (en)
INCLUDES	organic gardening (en) organic husbandry (en)
PRODUCES	organic foods (en)
IN OTHER LANGUAGES	زراعة عضوية — Arabic 有机农业 — Chinese 有机耕作

LiveSeeding

Ontologies

Navigation

TE=Term, TR=Trait, ME=Method, and SC=Scale

- TE Wheat traits
  - TE Abiotic stress
  - TE Agronomic
    - TR Aboveground biomass at maturity
    - TR Agronomic score
    - TR Anther extrusion
    - TR Biomass production rate
      - ME BMPR Computation
        - SC g/m2/day
    - TR Crop ground cover
    - TR Forage dry matter
    - TR Grain moisture content
    - TR Grain number
    - TR Grain number per spike
    - TR Grain number per spikelet
    - TR Grain number per tiller

Concept details

Trait: Biomass production rate
Method: BMPR Computation
Scale: g/m2/day

A variable is the combination between a trait a method and a scale

Key	Value
variable_id	CO_321:0001601
variable_name	BMPR_Calc_gm2day
ontology_id	CO_321
ontology_name	Wheat
context_of_use	Nursery/Trial evaluation
institution	CIMMYT
scientist	Gemma Molero, Rosemary Shrestha, Julian Pietragalla

Variables

GYPR_Calc_gm2day | CGR_Calc_gm2day | CropGrwRate_Calc_gm2day | BMPR_Calc_gm2day

LiveSeeding

Standards and metadata

LiveSeeding

What is a metadata ?

What is Metadata?

Metadata is: Data 'reporting'

- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** were the data created?
- **WHERE** is it geographically?
- **HOW** were the data developed?
- **WHY** were the data developed?

Photo by Michelle Chang. All Rights Reserved

LiveSeeding

Why a standard ?

Why use a standard ?

- Analyze, compare, exchange data
- Publish datasets in international data warehouse

And what about a metadata standard ?

- Describe data richly and accurately, using the same vocabulary as the rest of the scientific community
- To make your metadata interoperable and enable other systems to use them

LiveSeeding

Three text formats frequently used for metadata

CSV

~~XLS~~

But xlsx is XML, so it's ok

XML

eXtensible Markup Language

JSON

JavaScript Object Notation

Comma Separated Values

```
Sample_ alias, date, source
A, 20200802, blood
B, 20200802, feces
C, 20200802, skin
```

Human readable

Mostly machine readable —>

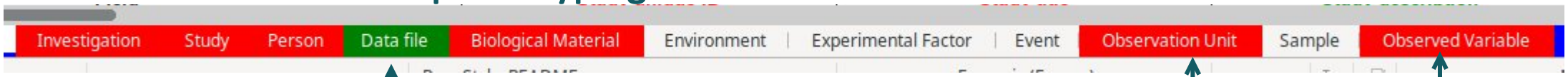
```
<SAMPLE_SET>
          <SAMPLE alias="A">

          <date>20200802</date>
          <source>blood</source>
     </SAMPLE>
          <SAMPLE alias="B">

          <date>20200802</date>
          <source>feces</source>
     </SAMPLE>
          <SAMPLE alias="C">

          <date>20200802</date>
          <source>skin</source>
     </SAMPLE>
</SAMPLE_SET>
```

```
"SAMPLE_SET": {
          "SAMPLE":[
     {
          "alias": "A",

          "date":"20200802",
          "source":"blood"
     },
          {
          "alias": "B",

          "date":"20200802",
          "source":"feces"
     },
  {
          "alias": "C",

          "date":"20200802",
          "source":"skin"
     }
     ]
  }
}
```

LiveSeeding

Minimum Information for Plant Phenotyping Experiments (MIAPPE)

MIAPPE is a standard for phenotyping metadata



The object that you will describe :
often a biological material at a location (plot)

Variables that you will measure such as a combination between a trait x a method x a scale

tidy data file :
ObservationUnit x ObservationVariables

=> tidy data file + MIAPPE metadata file is a FAIR dataset !

LiveSeeding

miappe

https://www.miappe.org/

Include established standard

MultiCrop Passport Descriptors (MCPD): genetic resources description

Crop Ontology: descriptions of methods, traits, measures

Licenses

LiveSeeding

Why a license

Give a framework for data sharing and reuse

- Allows users to be granted certain usage rights
- May include restrictions on use
- Strongly recommended in all cases to clearly display the related rights

Recommended Licenses

- Widely used license
- Compliant with other existing licenses (easy data aggregation)
- Taking into account the potential of the data and the restrictions applied (Etalab for distribution in France, Creative Commons for international distribution)

LiveSeeding

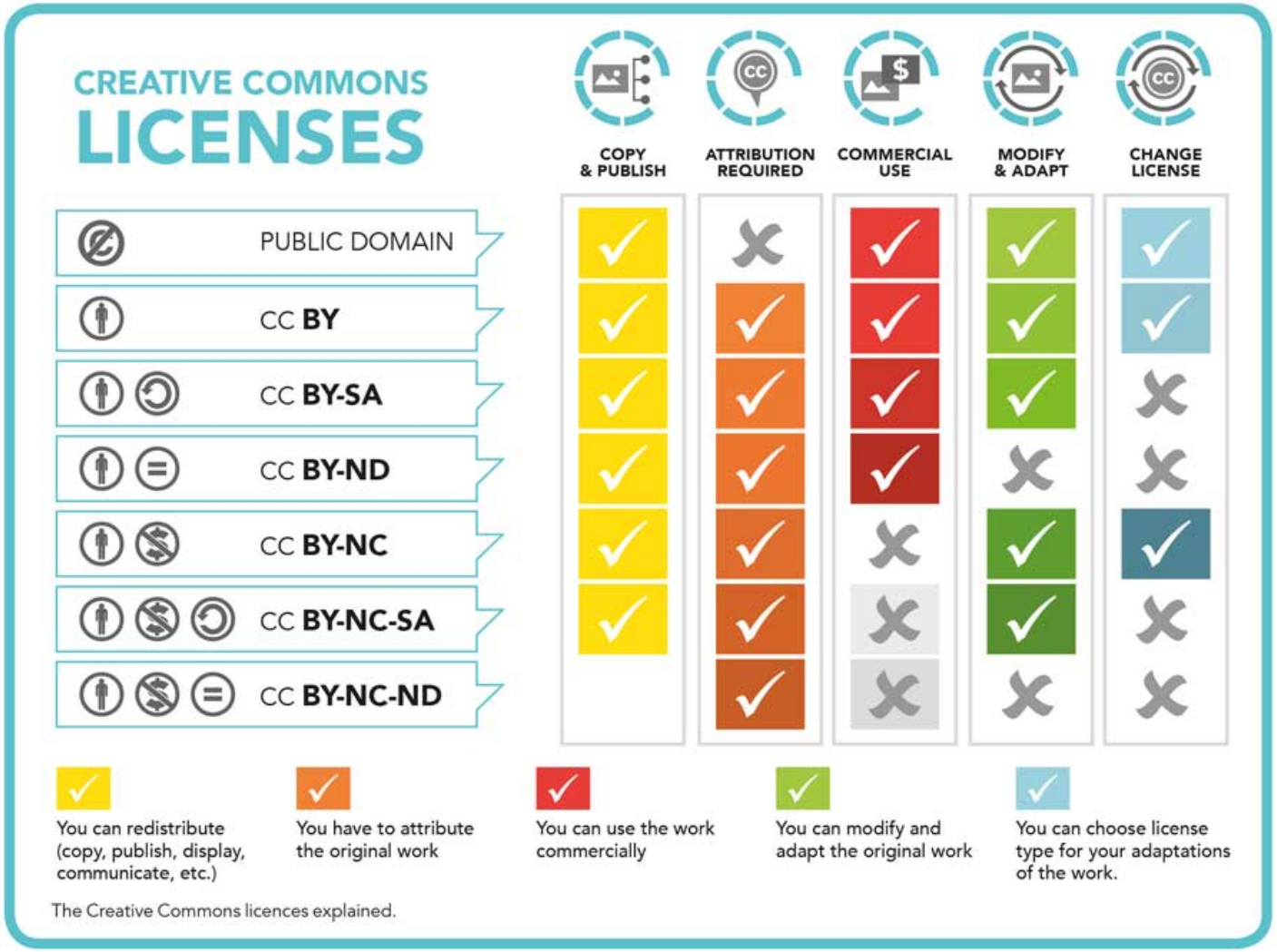
Creative Commons licenses

CC-0 : public domain, not cited

ND : impossible to modify/adapt data including aggregation

NC : not recommended because you prevent any commercial use, including a non-profit action such as a training that would need inscription fee

CC BY and CC BY-SA are the mostly recommended licenses



LiveSeeding

Code licenses

Permissive license means the license of the modified code can be changed. You have to be cited but the resulting code be proprietary.

Copyleft license means that the modified code must published under the same term of license.

We talk also about “viral” license.

All those licenses are good to use regarding the context.

	APACHE	BSD	MIT	GPL v3 Free as in Freedom	LGPL v3 Free as in Freedom	AGPL v3 Free as in Freedom
Type	Permissive	Permissive	Permissive	Copyleft	Copyleft	Copyleft
Provides copyright protection	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE
Can be used in commercial applications	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE	✓ TRUE
Provides an explicit patent license	✓ TRUE	✗ FALSE	✗ FALSE	✗ FALSE	✗ FALSE	✗ FALSE
Can be used in proprietary (closed source) projects	✓ TRUE	✓ TRUE	✓ TRUE	✗ FALSE	✗ FALSE partially	✗ FALSE for web
Popular open-source and free projects	Kubernetes Swift Firebase	Django React Flutter	Angular.js JQuery, .NET Core Laravel	Joomla Notepad++ MySQL	Qt SharpDevelop	SugarCRM Launchpad

LiveSeeding

Data repository

LiveSeeding

What kind of data warehouse

Thematic data warehouses

+

- Accurate dataset description and metadata
- Good quality

-

- Do not exists for all data types (phenotyping)
- Publication/curation of data time consuming

General data warehouse

+

- All data types
- Minimal metadata to publish
- Fast submission

-

- Poor description
- non-standard format
- Less FAIR

LiveSeeding

Examples of data warehouses

Thematic

open

GBIF
PANGAEA. Data Publisher for Earth & Environmental Science
NCBI
ENA European Nucleotide Archive
GEO Gene Expression Omnibus
EMBL-EBI UniProt
EMBL-EBI
BioFresh
ArrayExpress
BioStudies.
CELL IMAGE LIBRARY

controlled access

EUROPEAN GENOME-PHENOME ARCHIVE

LiveSeeding

General

zenodo
B2SHARE
EUDAT
DRYAD

Editor

Mendeley Data
(GIGA)^{n}DB

Organizations

Data INRAE
GnpIS
e!DAL Plant Genomics & Phenomics Research Data Repository

National (France)

RÉPUBLIQUE FRANÇAISE Liberté Égalité Fraternité
recherche.data.gouv.fr

Community

maizeGDB Maize Genetics and Genomics Database
BREEDBASE
Sol Genomics Network

Tips to choose your data warehouse

Consider these criterion to choose you data warehouse :

- Choose a thematic warehouse especially if it is an established one such as ENA (European Nucleotide Archive) for sequence data
- Choose a national/institutional warehouse if you have one in your country
- If you don’t find a warehouse that fits this criteria find one this is : open source, general scope, provide a DOI, robustly funded.

LiveSeeding

Short test

Download the quiz :

https://tinyurl.com/96dfw6wp

And send it to yannick.de-oliveira@inrae.fr

LiveSeeding

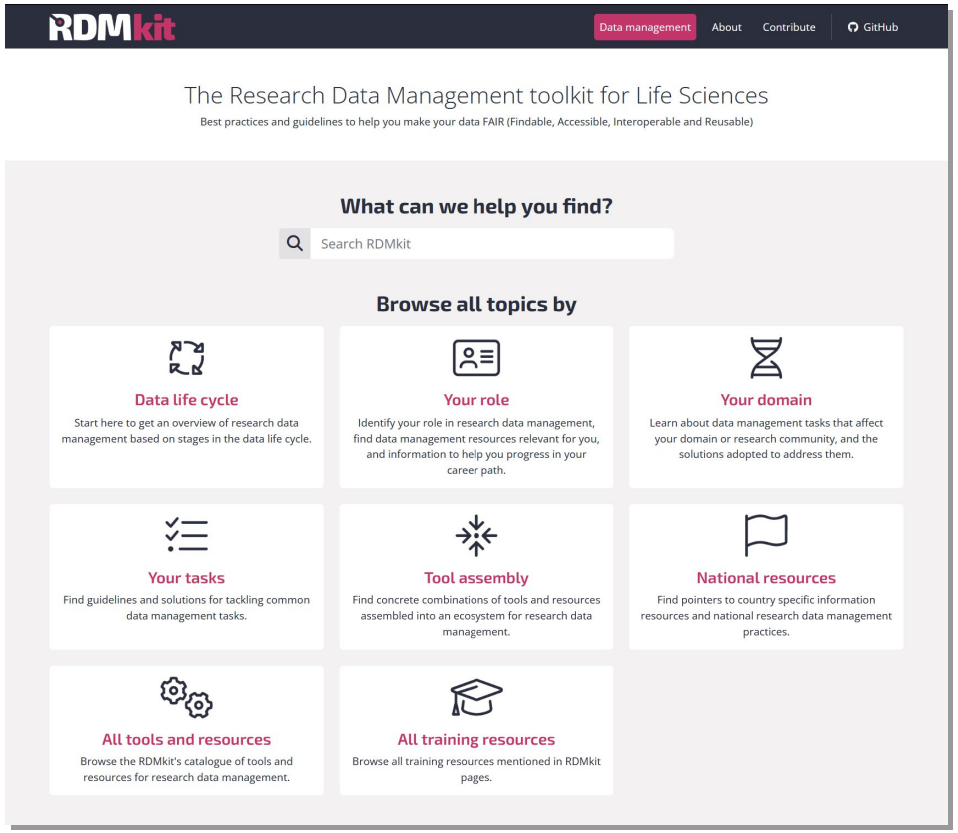
Key messages

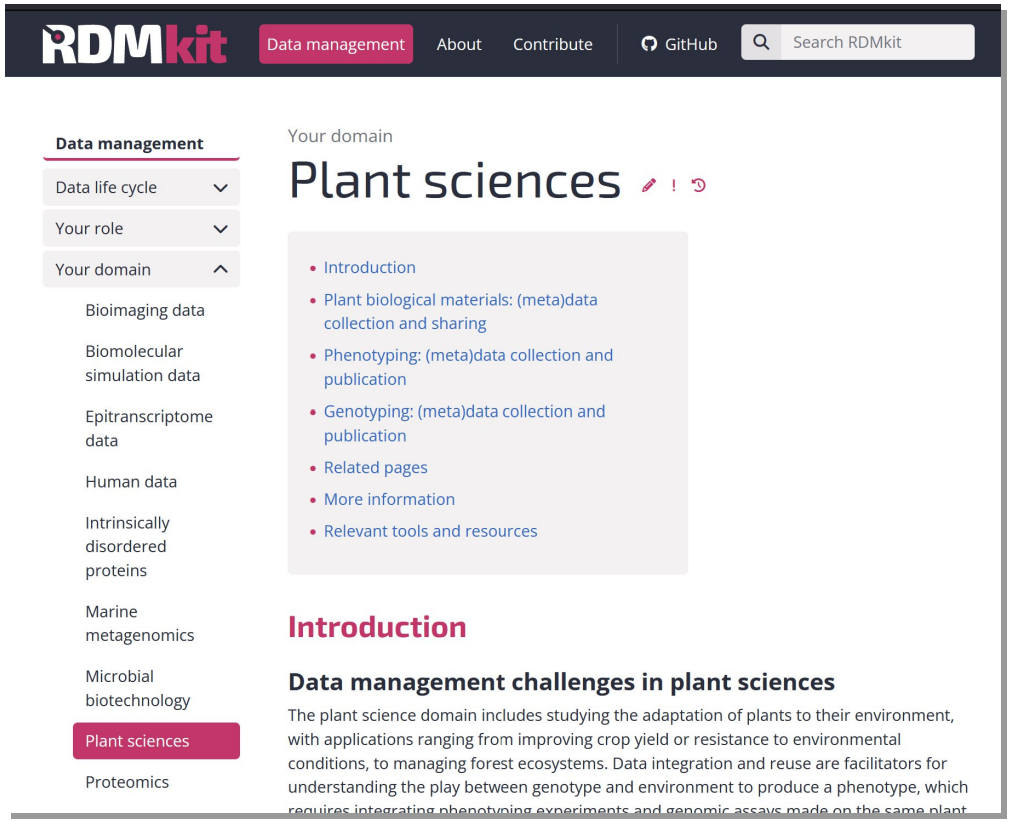
- Be FAIR with your data :)
- Use tidy data, standard and controlled vocabularies to increase quality of your data
- Share your data with appropriate license and warehouse to make them reusable

LiveSeeding

Resources

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





LiveSeeding

Resources

- **Controlled vocabularies :**

  https://agrovoc.fao.org/browse/agrovoc/en/

  https://cropontology.org/

- **Data Management Plan :**

  https://dmp.opidor.fr/

  https://ds-wizard.org

- **Metadata standard :**

  https://www.miappe.org/

  https://fairsharing.org/

- **Interoperability**

  https://brapi.org/

- **Licenses**

  https://creativecommons.org/

- **Data warehouse**

  https://zenodo.org/

  https://recherche.data.gouv.fr/fr

LiveSeeding

Module 2 – Unit 3 Additional available materials

1.

LiveSeeding

Initial steps toward reproducible research

Karl Broman's tutorial: http://kbroman.org/steps2rr/

1. *Everything with a script:*
   - *Get the data in the most-raw form, Download external data, Convert a data file, Don't hand-edit data files, Data cleaning should be in scripts, Analysis should be in scripts, Save your seeds for random number generation.*
2. *Organize your data and code:*
   - *Encapsulate everything within one directory, Separate raw data from derived data, Separate the data from the code, Use relative paths, Choose file names carefully, Avoid using "final" in a file name (rather version number), Write ReadMe files*
3. *Automate the process*
4. *Turn scripts into reproducible reports*
5. *Turn repeated code into functions*
6. *Package functions for reuse*

Module 2 – Unit 3 Additional available materials

Initial steps toward reproducible research:

- Karl Broman's tutorial: http://kbroman.org/steps2rr/

- Teach yourself:

- *Software Carpentry (https://software-carpentry.org/),*
- *MOOC RR (https://www.fun-mooc.fr/fr/cours/recherche-reproductible-principes-methodologiques-pour-une-science-transparente/ in French),*
- *Forum RR (https://forum.recherche-reproductible.fr/)*

LiveSeeding

LiveSeeding

LiveSeeding

Thank you !