ado2019 : journée thématique sur les autorités de données

8 avr. 2019 Toulouse (France)

FAIR principles, a new opportunity to improve the data lifecycle

Preliminary question :

Data Capture ?

Daniel Jacob

INRA UMR 1332 BFP – Metabolism Group

Bordeaux Metabolome Facility

INRA SCIENCE & IMPACT

Is a

transversal INRA CATI(*) whose vocation is to

promote, support and valorize

the implementation of

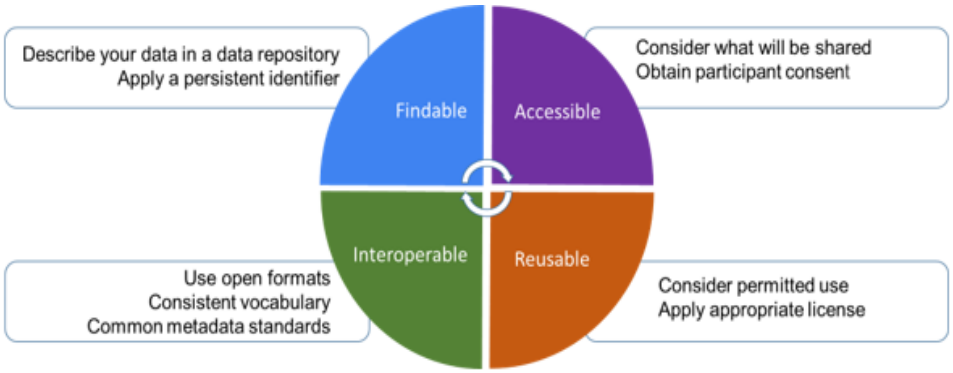
data and code management and sharing

for the purpose of

reuse and reproducibility of science.

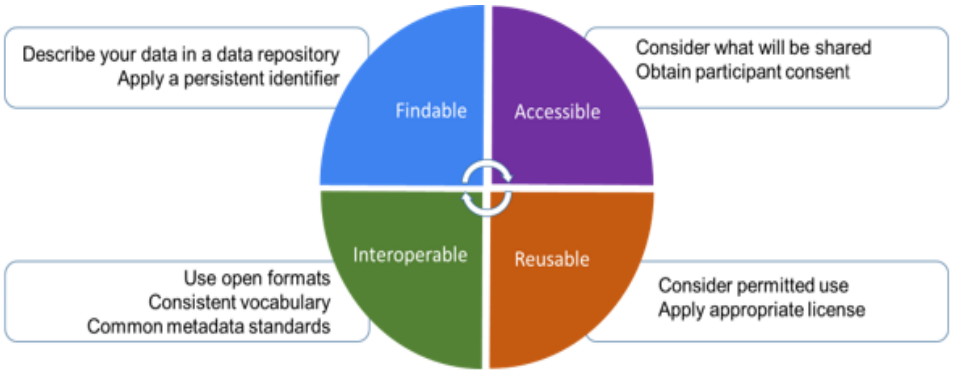
(*) *Centre Automatisé de Traitement de l'Information* - Automated Information Processing Center

THE FAIR DATA PRINCIPLES



The "FAIR" principles **define the basis** for data sharing easily to find, accessible, interoperable and reusable.

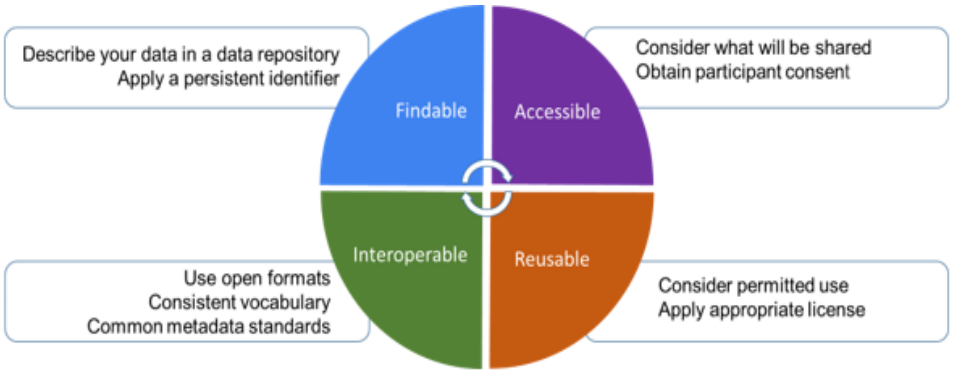
THE FAIR DATA PRINCIPLES



The "FAIR" principles **define the basis** for data sharing easily to find, accessible, interoperable and reusable.

However, it is to the communities to specify the actions necessary for their implementation

THE FAIR DATA PRINCIPLES



The "FAIR" principles **define the basis** for data sharing easily to find, accessible, interoperable and reusable.

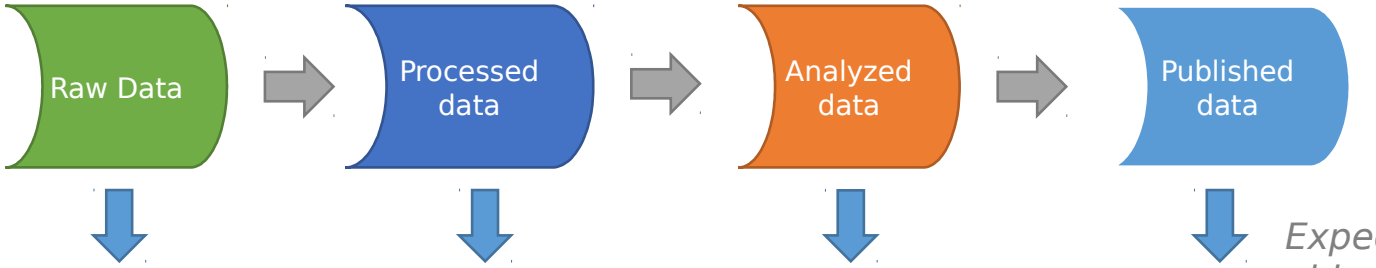
However, it is to the communities to specify the actions necessary for their implementation

The implementation of FAIR principles is a process that must be thought of in a progressive and community-oriented way.

It must be integrated into existing practices to ensure that they evolve without interruption and in a way that is acceptable to the various actors

Data flow

Studies with a design of experiment (DoE)



Specialized Data Repositories

EMBL-EBI
metabolomeXchange
ProteomeXchange
EUROPEAN GENOME-PHENOME ARCHIVE

Processed data is the raw data processed in a way so that they **can highlight some features**, i.e. some type of variables **linked to the focus of the study**.

Expected objectives

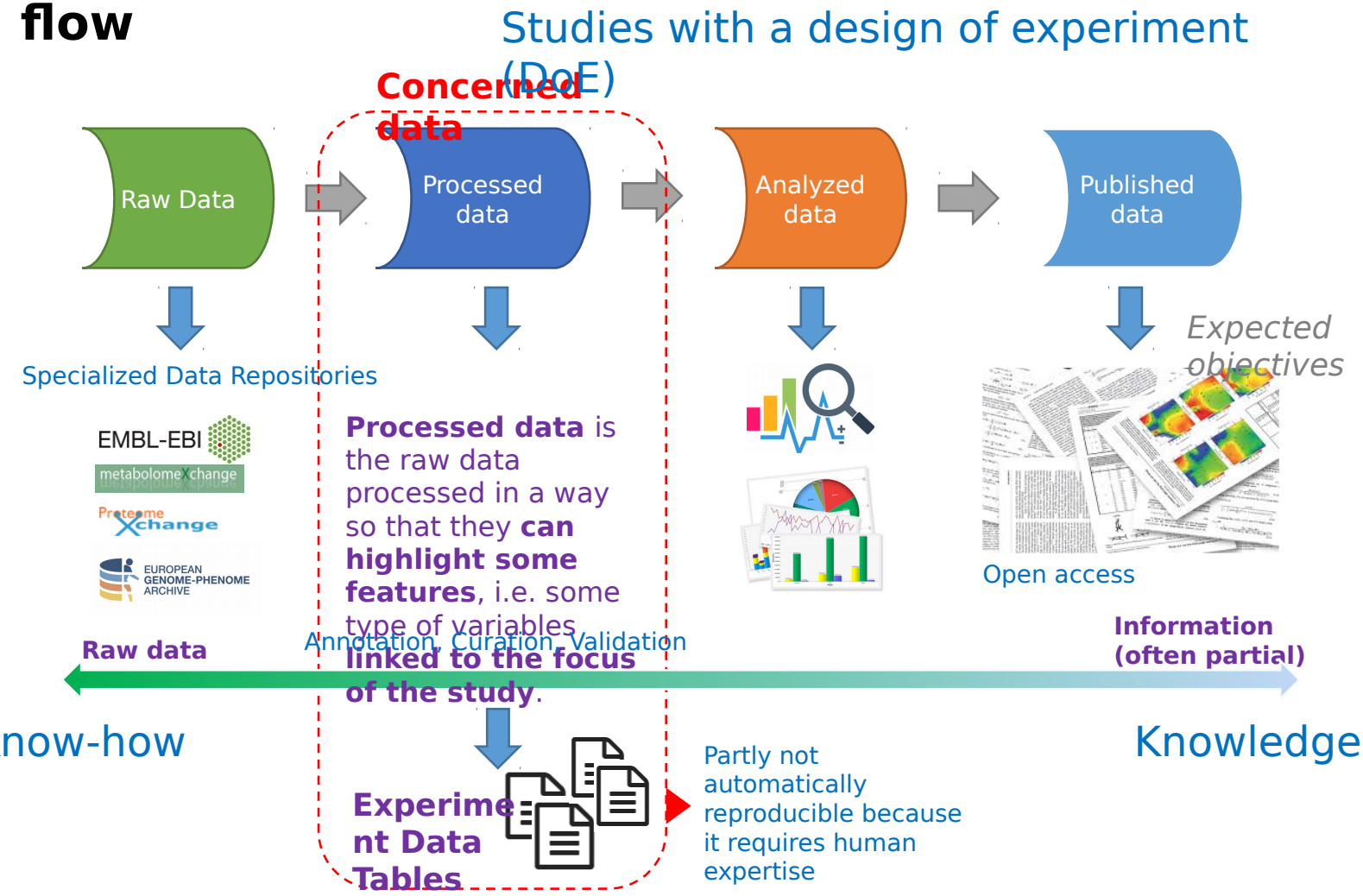
Open access

Annotation, Curation, Validation

Raw data ← → **Information (often partial)**

Know-how — Knowledge

Data flow


Studies with a design of experiment (DoE)
Concerned data
Raw Data
Processed data
Analyzed data
Published data
Specialized Data Repositories
EMBL-EBI
metabolomeXchange
ProteomeXchange
EUROPEAN GENOME-PHENOME ARCHIVE
Processed data is the raw data processed in a way so that they can highlight some features, i.e. some type of variables linked to the focus of the study.
Expected objectives
Open access
Raw data
Annotation, Curation, Validation
Information (often partial)
Know-how
Knowledge
Experiment Data Tables
Partly not automatically reproducible because it requires human expertise

Data flow

Studies with a design of experiment (DoE)

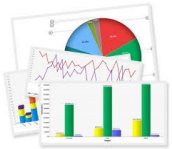
Concerned data

Raw Data → Processed data → Analyzed data → Published data

Specialized Data Repositories

EMBL-EBI
metabolomeXchange
ProteomeXchange
EUROPEAN GENOME-PHENOME ARCHIVE

Processed data is the raw data processed in a way so that they **can highlight some features**, i.e. some type of variables **linked to the focus of the study**.



Expected objectives

Open access

Raw data ← Annotation, Curation, Validation → **Information (often partial)**

Know-how — Knowledge

Experiment Data Tables

Partly not automatically reproducible because it requires human expertise

⇨ Data capture …

as far upstream as possible

The "data lifecycle" must therefore be integrated into the scientific research process

Data capture... **as far upstream as possible**

Take into account users operating methods and work habits

Data capture... **as far upstream as possible**

Take into account users operating methods and work habits

Spreadsheet as a central tool

Despite all their drawbacks

e.g. multiple information in a format without internal structure

This does not take away any of their benefits

Universal tool

Data capture... **as far upstream as possible**

Take into account users operating methods and work habits

Spreadsheet as a central tool

Despite all their drawbacks

e.g. multiple information in a format without internal structure

This does not take away any of their benefits

Universal tool

- **But: Repetitive and tedious tasks**

Gathering Data and Preparing Data :

- lot of data manipulation, mainly in the form of tables,
- combine data sets according to a common field (identifiers)

Modelisation :

- selection of subset of data then many repetitions of complex processing operations
- according to a very varied parametrization (scenarios).

Data capture... **as far upstream as possible**

Take into account users operating methods and work habits

Spreadsheet as a central tool

Despite all their drawbacks

e.g. multiple information in a format without internal structure

This does not take away any of their benefits

Universal tool

- **But: Repetitive and tedious tasks**

Gathering Data and Preparing Data :

- lot of data manipulation, mainly in the form of table,
- combine data sets according to a common field (identifiers)

Modelisation :

- selection of subset of data then many repetitions of complex processing operations
- according to a very varied parametrization (scenarios).

Handling of all these tasks related to data management

Provide services

Allow users to gain efficiency where they would like to gain efficiency

Data capture... **as far upstream as possible**

Take into account users operating methods and work habits

Spreadsheet as a central tool

Despite all their drawbacks

e.g. multiple information in a format without internal structure

This does not take away any of their benefits

Universal tool

But: Repetitive and tedious tasks

Gathering Data and Preparing Data :

- lot of data manipulation, mainly in the form of table,
- combine data sets according to a common field (identifiers)

Modelisation :

- selection of subset of data then many repetitions of complex processing operations
- according to a very varied parametrization (scenarios).

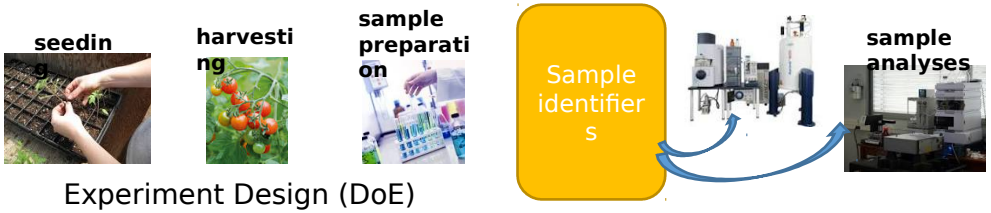
Handling of all these tasks related to data management

Promote good practices

Provide services

Allow users to gain efficiency where they would like to gain efficiency

romote good practices

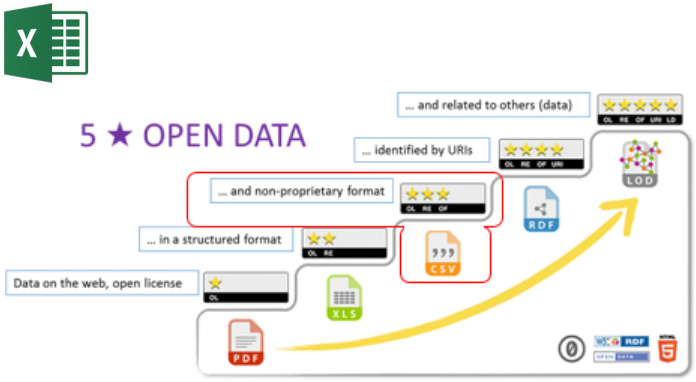


Experiment Design (DoE)

samples : Sample features

A	B	C	D	E	F	G	H	I	J
SampleID	Treatment	DevStage	FruitAge	FruitPosition	FruitDiamete	FruitHeight	FruitFW	Rank	Truss
115	Control	FF.01	07DPA	3	11.95	10.42	0.81	A	T7
121	Control	FF.03	22DPA	3	36.13	31.77	21.43	A	T6
164	Control	FR.01	42DPA	2	51.09	46.85	64.05	A	T5
353	Control	FR.04	55DPA	5	48.28	43.35	66.64	A	T5
355	Control	FR.04	55DPA	3	49.84	44.93	66.98	A	T5
413	Control	FR.02	47DPA	1	60.48	54.23	106.13	A	T7
512	Control	FF.03	21DPA	NA	41	35.82	37.22	A	TA
117	Control	FF.01	07DPA	3	13.44	12.39	1.14	A	T7
536	Control	FR.02	47DPA	NA	59.4	49.05	87.28	A	TA
544	Control	FR.03	50DPA	NA	57.31	47.69	92.86	A	TA
158	Control	FF.04	35DPA	5	58.38	49.3	92.86	A	T5
109	Control	FF.03	22DPA	7	43.37	35.77	38.73	A	T5
134	Control	FF.02	15DPA	3	27.89	23.8	9.88	A	T7
31	Control	FF.01	08DPA	4	NA	NA	0.48	A	T6
179	Control	FF.03	28DPA	3	53.68	45.43	65.34	A	T7
383	Control	FF.04	34DPA	5	47.04	41.19	48.96	A	T7
425	Control	FR.04	55DPA	2	62.74	50.27	115.3	A	T7
520	Control	FF.03	30DPA	NA	48.86	41.52	52.94	A	TA
419	Control	FR.03	50DPA	2	55.63	48.02	86.79	A	T7
138	Control	FF.02	15DPA	6	27.96	22.14	9.69	A	T7
143	Control	FF.03	29DPA	4	48.45	42.92	51.35	A	T6
365	Control	FR.02	47DPA	5	55.11	44.9	71.82	A	T6
127	Control	FF.03	27DPA	3	45.71	43.28	47.8	A	T5
188	Control	FR.01	42DPA	3	55.38	47.1	77.39	A	T6

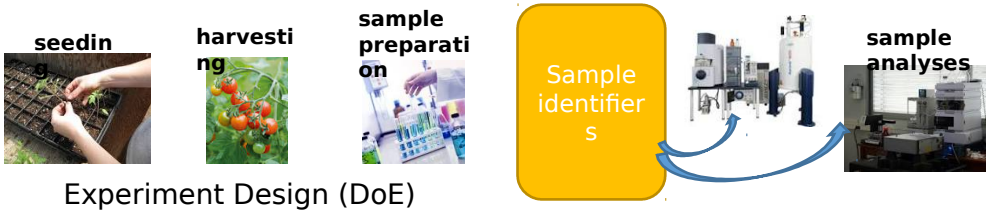
Data



necessary and indispensable step towards « **Linked Open Data** ».

Promote non-proprietary format like CSV or TSV

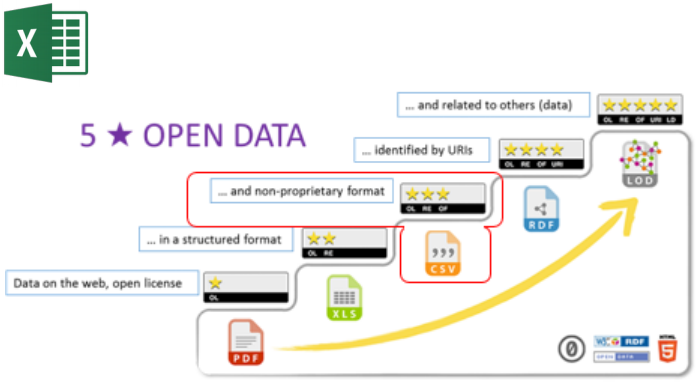
romote good practices



Experiment Design (DoE)

samples : Sample features

A	B	C	D	E	F	G	H	I	J
SampleID	Treatment	DevStage	FruitAge	FruitPosition	FruitDiamete	FruitHeight	FruitFW	Rank	Truss
115	Control	FF.01	07DPA	3	11.95	10.42	0.81	A	T7
121	Control	FF.03	22DPA	3	36.13	31.77	21.43	A	T6
164	Control	FR.01	42DPA	2	51.09	46.85	64.05	A	T5
353	Control	FR.04	55DPA	5	48.28	43.35	66.64	A	T5
355	Control	FR.04	55DPA	3	49.84	44.93	66.98	A	T5
413	Control	FR.02	47DPA	1	60.48	54.23	106.13	A	T7
512	Control	FF.03	21DPA	NA	41	35.82	37.22	A	TA
117	Control	FF.01	07DPA	3	13.44	12.39	1.14	A	T7
536	Control	FR.02	47DPA	NA	59.4	49.05	87.28	A	TA
544	Control	FR.03	50DPA	NA	57.31	47.69	92.86	A	TA
158	Control	FF.04	35DPA	5	58.38	49.3	92.86	A	T5
109	Control	FF.03	22DPA	7	43.37	35.77	38.73	A	T5
134	Control	FF.02	15DPA	3	27.89	23.8	9.88	A	T7
31	Control	FF.01	08DPA	4	NA	NA	0.48	A	T6
179	Control	FF.03	28DPA	3	53.68	45.43	65.34	A	T7
383	Control	FF.04	34DPA	5	47.04	41.19	48.96	A	T7
425	Control	FR.04	55DPA	2	62.74	50.27	115.3	A	T7
520	Control	FF.03	30DPA	NA	48.86	41.52	52.94	A	TA
419	Control	FR.03	50DPA	2	55.63	48.02	86.79	A	T7
138	Control	FF.02	15DPA	6	27.96	22.14	9.69	A	T7
143	Control	FF.03	29DPA	4	48.45	42.92	51.35	A	T6
365	Control	FR.02	47DPA	5	55.11	44.9	71.82	A	T6
127	Control	FF.03	27DPA	3	45.71	43.28	47.8	A	T5
188	Control	FR.01	42DPA	3	55.38	47.1	77.39	A	T6



necessary and indispensable step towards « **Linked Open Data** ».

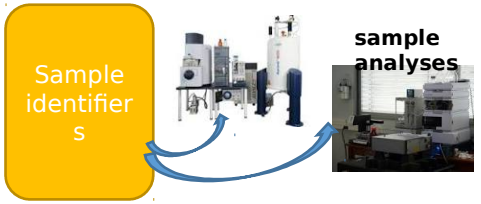
Promote non-proprietary format like CSV or TSV

Data

Well organized data

- Each variable forms a column
- Each observation forms a line
- Each "observational entity" forms a table

romote good practices



Experiment Design (DoE)

samples : Sample features

A	B	C	D	E	F	G	H	I	J
SampleID	Treatment	DevStage	FruitAge	FruitPosition	FruitDiamete	FruitHeight	FruitFW	Rank	Truss
115	Control	FF.01	07DPA	3	11.95	10.42	0.81	A	T7
121	Control	FF.03	22DPA	3	36.13	31.77	21.43	A	T6
164	Control	FR.01	42DPA	2	51.09	46.85	64.05	A	T5
[illegible]	[illegible]	[illegible]	55DPA	5	48.28	43.35	66.64	A	T5
[illegible]	[illegible]	[illegible]	55DPA	3	49.84	44.93	66.98	A	T5
[illegible]	[illegible]	[illegible]	47DPA	1	60.48	54.23	106.13	A	T7
512	Control	FF.03	21DPA	NA	41	35.82	37.22	A	TA
117	Control	[illegible]	[illegible]	3	13.44	12.39	1.14	A	T7
536	Control	[illegible]	[illegible]	NA	59.4	49.05	87.28	A	TA
544	Control	[illegible]	[illegible]	NA	57.31	47.69	92.86	A	TA
158	Control	FF.04	35DPA	[illegible]	[illegible]	[illegible]	2.86	A	T5
109	Control	FF.03	22DPA	[illegible]	[illegible]	[illegible]	3.73	A	T5
134	Control	FF.02	15DPA	[illegible]	[illegible]	[illegible]	3.88	A	T7
31	Control	FF.01	08DPA	4	NA	NA	0.48	A	T6
179	Control	FF.03	28DPA	3	53.68	45.43	65.34	A	T7
383	Control	FF.04	34DPA	5	47.04	41.19	[illegible]	[illegible]	[illegible]
425	Control	FR.04	55DPA	2	62.74	50.27	1	[illegible]	[illegible]
520	Control	FF.03	30DPA	NA	48.86	41.52	5	[illegible]	[illegible]
419	Control	FR.03	50DPA	2	55.63	48.02	[illegible]	A	[illegible]
138	Control	FF.02	15DPA	6	27.96	22.14	9.69	A	T7
143	Control	FF.03	29DPA	4	48.45	42.92	51.35	A	T6
365	Control	FR.02	47DPA	5	55.11	44.9	71.82	A	T6
127	Control	FF.03	27DPA	3	45.71	43.28	47.8	A	T5
188	Control	FR.01	42DPA	3	55.38	47.1	77.39	A	T6

Identifier

Factors

Quantitatives

Qualitatives

Data

Well organized data

- Each variable forms a column
- Each observation forms a line
- Each "observational entity" forms a table

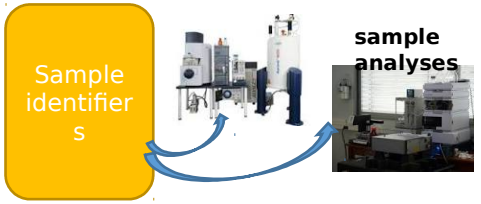
Whatever the kind of experiment, this assumes a design of experiment (**DoE**) involving individuals, samples or whatever things, as the main objects of study (e.g. plants, animal, bacteria, tissues, ...)

This also assumes the observation of dependent variables resulting of effects of some controlled experiment factors.

Moreover, the objects of study have usually an identifier for each of them, and the variables can be quantitative or qualitative.

Promote non-proprietary format like CSV or TSV

romote good practices



Experiment Design (DoE)

samples : Sample features

SampleID	Treatment	DevStage	FruitAge	FruitPosition	FruitDiamete	FruitHeight	FruitFW	Rank	Truss
115	Control	FF.01	07DPA	3	11.95	10.42	0.81	A	T7
121	Control	FF.03	22DPA	3	36.13	31.77	21.43	A	T6
164	Control	FR.01	42DPA	2	51.09	46.85	64.05	A	T5
			55DPA	5	48.28	43.35	66.64	A	T5
			55DPA	3	49.84	44.93	66.98	A	T5
			47DPA	1					
512	Control	FF.03	21DPA	NA					
117	Control			3					
536	Control			NA					
544	Control			NA					
158	Control	FF.04	35DPA						
109	Control	FF.03	22DPA						
134	Control	FF.02	15DPA						
31	Control	FF.01	08DPA	4					
179	Control	FF.03	28DPA	3					
383	Control	FF.04	34DPA	5					
425	Control	FR.04	55DPA	2					
520	Control	FF.03	30DPA	NA					
419	Control	FR.03	50DPA	2					
138	Control	FF.02	15DPA	6	27.96	22.14	9.69	A	T7
143	Control	FF.03	29DPA	4	48.45	42.92	51.35	A	T6
365	Control	FR.02	47DPA	5	55.11	44.9	71.82	A	T6
127	Control	FF.03	27DPA	3	45.71	43.28	47.8	A	T5
188	Control	FR.01	42DPA	3	55.38	47.1	77.39	A	T6

Identifier

Factors

Qua

Data

Description of the different columns within data files

Shortname	Description		
SampleID	Pool of several harvests		Identifier
Treatment	Treatment applied on plants		Factor
DevStage	fruit development stage		Factor
FruitAge	fruit age	Days post-anthesis (dpa)	Factor
FruitDiameter	Fruit diameter	mm	Variable
FruitHeight	Fruit height	mm	Variable
FruitFW	Fruit Fresh Weight(g)	g	Variable
Rank	Row of the invidual plant on the table		Feature
Truss	Position on the stem of the truss		Feature

Metadata

categories
- identifier
- factor
- qualitative
- quantitative

⇒ Metadata : not just on the "top" linked to datasets but **more deeply linked to the variables**.

Promote non-proprietary format like CSV or TSV

romote good practices Minimal but relevant Metadata

Metadata file allowing to associate a key concept to each **data subset file**

	A	B	C	D	E	F	G	H
1	rank	obtainedFrom	subset	identifier	file	description	CV_term_id	CV_term_name
2	1	0	plants	PlanteID	plants.csv	Plant features	http://purl.obolibrary.org/obo/PO_0000003	whole plant
3	2	1	harvests	Lot	harvests.csv	Harvest features	http://purl.obolibrary.org/obo/OBI_1110046	organ harvesting
4	3	2	samples	SampleID	samples.csv	Samples features	http://purl.obolibrary.org/obo/PO_0009001	fruit
5	4	3	compounds	SampleID	compounds.csv	Compound quantifications	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
6	5	3	enzymes	SampleID	enzymes.csv	Enzyme Features	http://purl.obolibrary.org/obo/OBI_0000427	enzyme

...

Entity ⇔ Observational entity (e.g. samples, compounds, ...)

Metadata file allowing each attribute (variable) to be annotated

	A	B	C	D	E	F	G	H
1	subset	attribute	entry	category	type	description	CV_term_id	CV_term_name
2	plants	PlanteID	plantid	identifier	numeric	Plant identifier	http://purl.obolibrary.org	individual organism identifier
3	plants	Row	row	qualitative	string	Row of the invidual plant on the table	http://ncicb.nci.nih.gov/x	Row
4	plants	Plant	plant		string	Code identifier of the individual plant	http://ncicb.nci.nih.gov/x	Discrete Set Coded String Data Type
5	plants	Treatment	treatment	factor	string	Treatment applied on plants	http://www.ebi.ac.uk/efo	environmental factor
6	plants	Genotype		qualitative	string	Genotype		
7	harvests	Lot	lot	identifier	numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
8	harvests	PlanteID			numeric	Plant identifier	http://purl.obolibrary.org	individual organism identifier
9	harvests	Truss		qualitative	string	Position on the stem of the truss	http://purl.obolibrary.org	stem node
10	harvests	HarvestDate			string	Harvest date		
11	harvests	HarvestHour			string	Harvest hour		
12	harvests	FruitAge	age	factor	string	fruit development stage	http://purl.obolibrary.org	fruit development stage
13	harvests	FruitPosition		qualitative	numeric	Poistion on the truss of the fruit	http://ncicb.nci.nih.gov/x	Position Number
14	harvests	FruitDiameter		quantitative	numeric	Fruit diameter (mm)	http://ncicb.nci.nih.gov/x	Diameter
15	harvests	FruitHeight		quantitative	numeric	Fruit height (mm)	http://ncicb.nci.nih.gov/x	Height
16	harvests	FruitFW		quantitative	numeric	Fruit Fresh Weight(g)	http://ncicb.nci.nih.gov/x	Weight
17	samples	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org	centrally registered identifier
18	samples	Lot			numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
19	samples	NbFruit			numeric	Fruit Number per sample		
20	samples	GellyFW		quantitative	numeric	Gelly Fred Weight(g) per sample		
21	samples	GellyFruit		quantitative	numeric	Gelly per Fruit (estimated g)		
22	samples	BER			string	BER		
23	compounds	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org	centrally registered identifier
24	compounds	DPA		factor	numeric	Day Per Anthesis		
25	compounds	MassBefore		quantitative	numeric	m av.extraction (g)		
26	compounds	MassMIA		quantitative	numeric	masse MIA (g)		
27	compounds	RDT		quantitative	numeric	Rdt (% MIA/DW)		
28	compounds	Starch1		quantitative	numeric	Dosage amidon (%poids/MIA)	http://purl.obolibrary.org	starch
29	compounds	Starch2		quantitative	numeric	amidon (g/gDW)	http://purl.obolibrary.org	starch
30	compounds	RHAMNOSE		quantitative	numeric	RHAMNOSE	http://purl.obolibrary.org	rhamnose

categories

- identifier
- factor
- qualitative
- quantitative

...

Attribute ⇔ Variable, Feature, ... (e.g. Plants, Fruits, Glucose, Rank, ...)

romote good practices Minimal but relevant Metadata

Metadata file allowing to associate a key concept to each **data subset file**

	A	B	C	D	E	F	G	H
1	rank	obtainedFrom	subset	identifier	file	description	CV_term_id	CV_term_name
2	1	0	plants	PlanteID	plants.csv	Plant features	http://purl.obolibrary.org/obo/PO_0000003	whole plant
3	2	1	harvests	Lot	harvests.csv	Harvest features	http://purl.obolibrary.org/obo/OBI_1110046	organ harvesting
4	3	2	samples	SampleID	samples.csv	Samples features	http://purl.obolibrary.org/obo/PO_0009001	fruit
5	4	3	compounds	SampleID	compounds.csv	Compound quantifications	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
6	5	3	enzymes	SampleID	enzymes.csv	Enzyme Features	http://purl.obolibrary.org/obo/OBI_0000427	enzyme

Optional

...

Entity ⇔ Observational entity (e.g. samples, compounds, ...)

Metadata file allowing each attribute (variable) to be annotated

	A	B	C	D	E	F	G	H
1	subset	attribute	entry	category	type	description	CV_term_id	CV_term_name
2	plants	PlanteID	plantid	identifier	numeric	Plant identifier	http://purl.obolibrary.org/	individual organism identifier
3	plants	Row	row	qualitative	string	Row of the invidual plant on the table	http://ncicb.nci.nih.gov/x	Row
4	plants	Plant	plant		string	Code identifier of the individual plant	http://ncicb.nci.nih.gov/x	Discrete Set Coded String Data Type
5	plants	Treatment	treatment	factor	string	Treatment applied on plants	http://www.ebi.ac.uk/efo	environmental factor
6	plants	Genotype		qualitative	string	Genotype		
7	harvests	Lot	lot	identifier	numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
8	harvests	PlanteID			numeric	Plant identifier	http://purl.obolibrary.org/	individual organism identifier
9	harvests	Truss		qualitative	string	Position on the stem of the truss	http://purl.obolibrary.org/	stem node
10	harvests	HarvestDate			string	Harvest date		
11	harvests	HarvestHour			string	Harvest hour		
12	harvests	FruitAge	age	factor	string	fruit development stage	http://purl.obolibrary.org/	fruit development stage
13	harvests	FruitPosition		qualitative	numeric	Poistion on the truss of the fruit	http://ncicb.nci.	
14	harvests	FruitDiameter		quantitative	numeric	Fruit diameter (mm)	http://ncicb.nci.	
15	harvests	FruitHeight		quantitative	numeric	Fruit height (mm)	http://ncicb.nci.	
16	harvests	FruitFW		quantitative	numeric	Fruit Fresh Weight(g)	http://ncicb.nci.nih.gov/x	Weight
17	samples	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org/	centrally registered identifier
18	samples	Lot			numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
19	samples	NbFruit			numeric	Fruit Number per sample		
20	samples	GellyFW		quantitative	numeric	Gelly Fred Weight(g) per sample		
21	samples	GellyFruit		quantitative	numeric	Gelly per Fruit (estimated g)		
22	samples	BER			string	BER		
23	compounds	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org/	centrally registered identifier
24	compounds	DPA		factor	numeric	Day Per Anthesis		
25	compounds	MassBefore		quantitative	numeric	m av.extraction (g)		
26	compounds	MassMIA		quantitative	numeric	masse MIA (g)		
27	compounds	RDT		quantitative	numeric	Rdt (% MIA/DW)		
28	compounds	Starch1		quantitative	numeric	Dosage amidon (%poids/MIA)	http://purl.obolibrary.org/	starch
29	compounds	Starch2		quantitative	numeric	amidon (g/gDW)	http://purl.obolibrary.org/	starch
30	compounds	RHAMNOSE		quantitative	numeric	RHAMNOSE	http://purl.obolibrary.org/	rhamnose

Optional

...

categories

- identifier
- factor
- qualitative
- quantitative

Attribute ⇔ Variable, Feature, ... (e.g. Plants, Fruits, Glucose, Rank, ...)

FAIR principles, a new opportunity to improve the data lifecycle

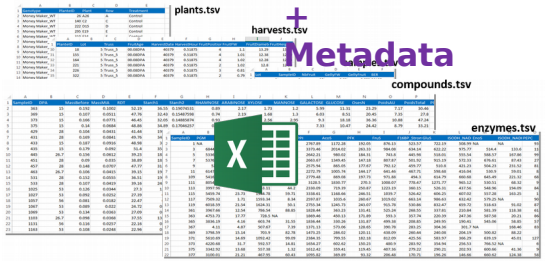
Provide services

Promote good practices

ODAM Framework

http://pmb-bordeaux.fr/dataexplorer/

Experiment Data Tables + Metadata



Additional data subsets can be added step by step, as soon as data are produced.

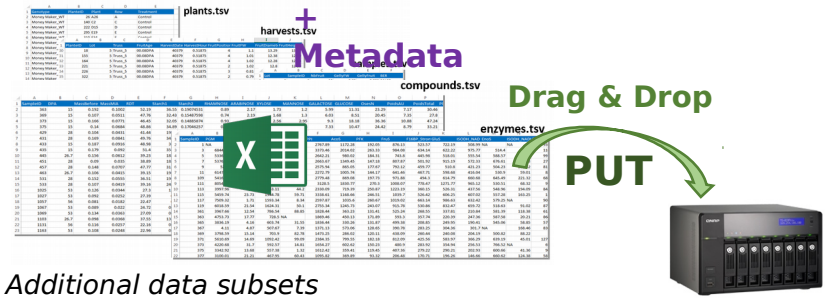
FAIR principles, a new opportunity to improve the data lifecycle

Provide services ← **Promote good practices**

ODAM Framework

http://pmb-bordeaux.fr/dataexplorer/

Experiment Data Tables + Metadata



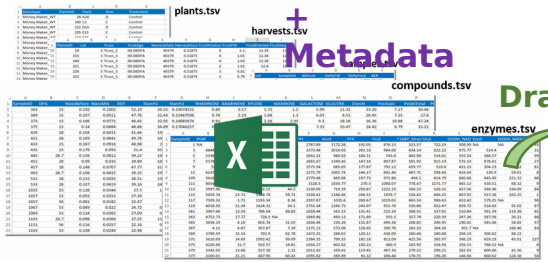
Additional data subsets can be added step by step, as soon as data are produced.

FAIR principles, a new opportunity to improve the data lifecycle

Provide services ← **Promote good practices**

ODAM Framework
http://pmb-bordeaux.fr/dataexplorer/

Experiment Data Tables + Metadata



Additional data subsets can be added step by step, as soon as data are produced.

Drag & Drop

PUT

Storage and sharing with all partners

Multiscale deployment
Local / Intranet / Internet

VM / docker

GET

Software layer transparent from the user's point of view

Data can be downloaded, explored and mined

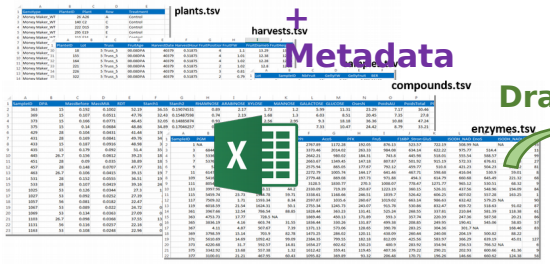
FAIR
INTEROPERABLE

FAIR principles, a new opportunity to improve the data lifecycle

Provide services ← **Promote good practices**

ODAM Framework
http://pmb-bordeaux.fr/dataexplorer/

Experiment Data Tables + Metadata



Additional data subsets can be added step by step, as soon as data are produced.

Drag & Drop

PUT

Storage and sharing with all partners

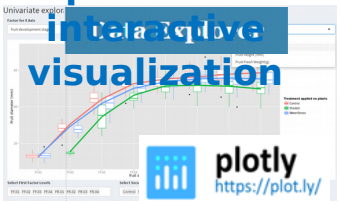
Multiscale deployment
Local / Intranet / Internet

VM / docker

GET

Data can be downloaded, explored and mined

Multi-criteria exploration & interactive visualization



Merging & selection of data subsets

Export

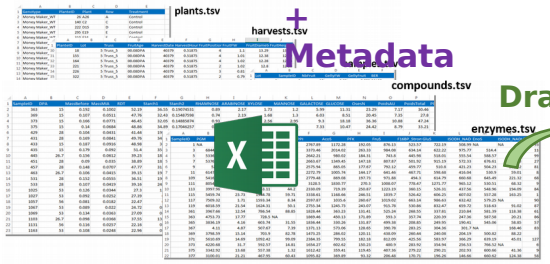
FAIR
I NTEROPERABLE

FAIR principles, a new opportunity to improve the data lifecycle

Provide services ← **Promote good practices**

ODAM Framework
http://pmb-bordeaux.fr/dataexplorer/

Experiment Data Tables + Metadata



Additional data subsets can be added step by step, as soon as data are produced.

Drag & Drop

PUT

Storage and sharing with all partners

Multiscale deployment
Local / Intranet / Internet

VM / docker

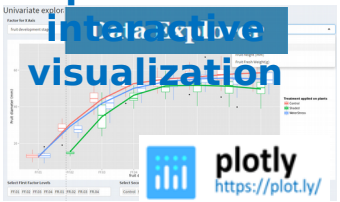
API

FAIR INTEROPERABLE

GET

Data can be downloaded, explored and mined

Multi-criteria exploration & interactive visualization



Merging & selection of data subsets

Export

Repetition of complex treatments according to very varied parameters

API

Develop if needed, lightweight tools

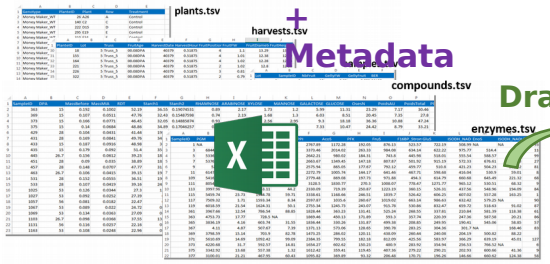
- R/Python scripts (Galaxy), lightweight GUI (R shiny)
- Python (Plotly Dash)

FAIR principles, a new opportunity to improve the data lifecycle

Provide services ← **Promote good practices**

ODAM Framework
http://pmb-bordeaux.fr/dataexplorer/

Experiment Data Tables + Metadata



Additional data subsets can be added step by step, as soon as data are produced.

Drag & Drop

PUT

Storage and sharing with all partners

Multiscale deployment
Local / Intranet / Internet

VM / docker

API

F
A
I NTEROPERABLE
R

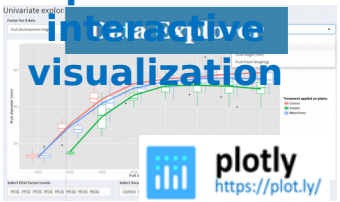
GET

Data can be downloaded, explored and mined

The Dataverse Project
https://data.inra.fr

Link

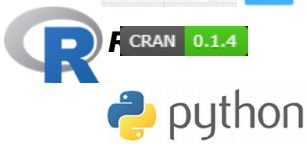
Multi-criteria exploration & interactive visualization



Merging & selection of data subsets

Export

Repetition of complex treatments according to very varied parameters



API

facilitate the subsequent data publication

F INDABLE
A CCESSIBLE
I
R E-USABLE

Develop if needed, lightweight tools

- R/Python scripts (Galaxy), lightweight GUI (R shiny)
- Python (Plotly Dash)

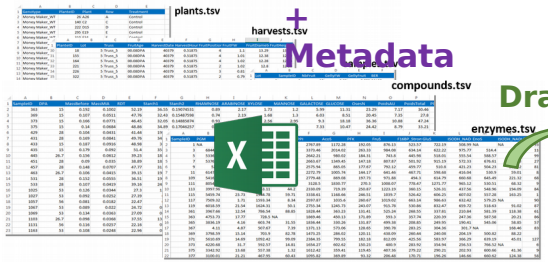
FAIR principles, a new opportunity to improve the data lifecycle

Promote good practices

Minimal effort

ODAM Framework
http://pmb-bordeaux.fr/dataexplorer/

Experiment Data Tables + Metadata



Additional data subsets can be added step by step, as soon as data are produced.

Drag & Drop

PUT

Storage and sharing with all partners

Multiscale deployment
Local / Intranet / Internet

VM / docker

API

F
A
I NTEROPERABLE
R

GET

Data can be downloaded, explored and mined

The Dataverse Project
https://data.inra.fr

F INDABLE
A CCESSIBLE
I
R E-USABLE

Link

Multi-criteria exploration & interactive visualization

API

facilitate the subsequent data publication

Provide services

Maximum efficiency

Merging & selection of data subsets

Export

Repetition of complex treatments according to very varied parameters

Develop if needed, lightweight tools

- R/Python scripts (Galaxy), lightweight GUI (R shiny)
- Python (Plotly Dash)

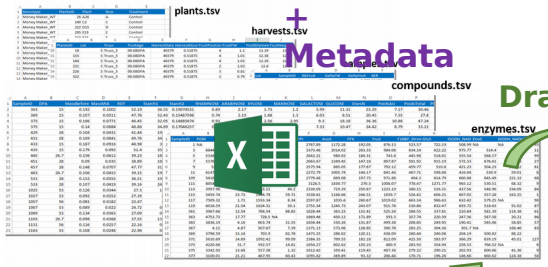
FAIR principles, a new opportunity to improve the data lifecycle

Data Lifecycle

ODAM Framework
http://pmb-bordeaux.fr/dataexplorer/

Data capture

Experiment Data Tables + Metadata



Additional data subsets can be added step by step, as soon as data are produced.

Data Backup

Drag & Drop
PUT

Storage and sharing with all partners

Multiscale deployment
Local / Intranet / Internet

FAIR
F
A
I NTEROPERABLE
R

GET

Data can be downloaded, explored and mined

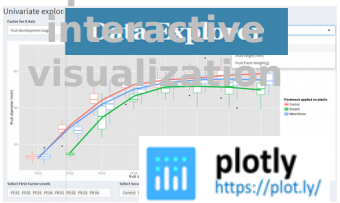
The Dataverse Project

FAIR
F INDABLE
A CCESSIBLE
I
R E-USABLE

Data Curation

Remove inconsistent values

Multi-criteria exploration & interactive visualization



Link

Merging & selection of data subsets

Export

Data Extension

Repetition of complex treatments according to very varied parameters

API

facilitate the subsequent data publication

Data Publication

https://data.inra.fr

The "data lifecycle" is thus integrated into the scientific research process

Develop if needed, lightweight tools
- R/Python scripts (Galaxy), lightweight GUI (R shiny)
- Python (Plotly Dash)

Opportunity for Reproducible Research

Reproduce the results presented in an article.

Fostered by the fact that data and metadata are all available and accessible by scripting languages (R, API)

Gathering Data and Preparing Data

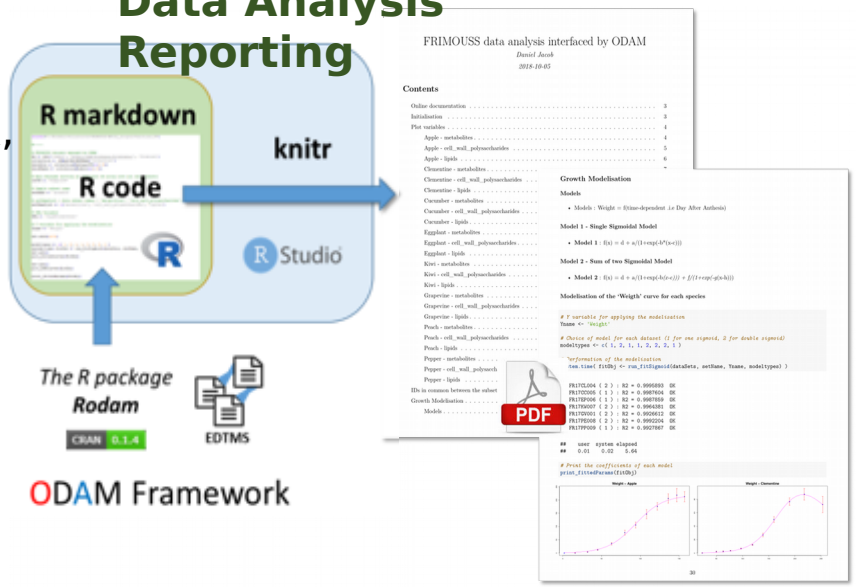
carried out by the ODAM approach

Merging & selection of data subsets

Repetition of complex treatments according to very varied parameters

F
A
I NTEROPERABLE
R

Data Analysis Reporting



Opportunity for Reproducible Research

Reproduce the results presented in an article.

Fostered by the fact that data and metadata are all available and accessible by scripting languages (R, API)

Gathering Data and Preparing Data

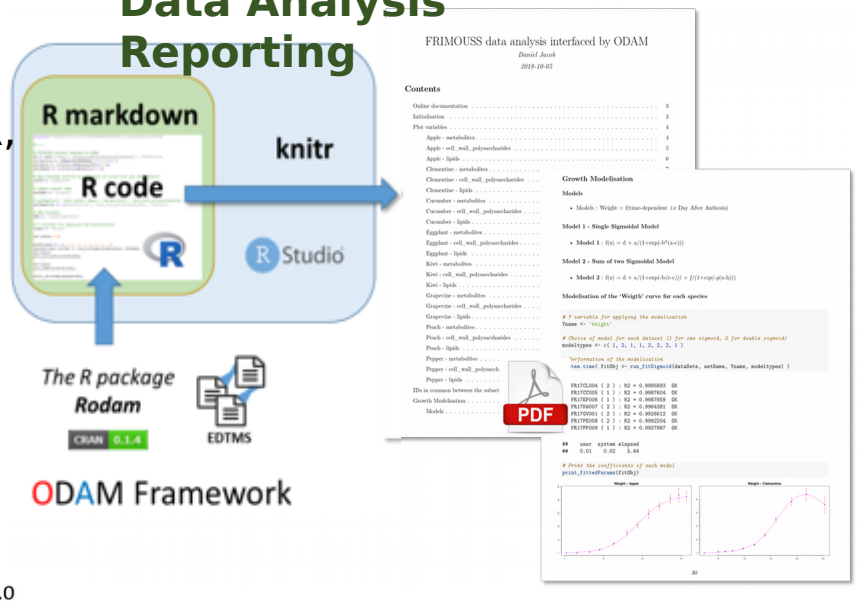
carried out by the ODAM approach

Merging & selection of data subsets

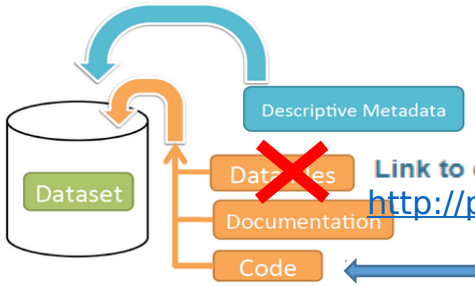
Repetition of complex treatments according to very varied parameters

F
A
I NTEROPERABLE
R

Data Analysis Reporting



Schematic Diagram of a **Dataset** in Dataverse 4.0



The Dataverse Project

https://data.inra.fr

Data Publication

Data Explorer

Link to data
http://pmb-bordeaux.fr/dataexplorer/

R scripts (Rmd)

F INDABLE
A CCESSIBLE
I
R E-
USABLE

Opportunity for the Web Of Data

Resource Description Framework (RDF)

Metadata

CV Term

Attributes

	A	B	C	D	E	F	G	H
1	subset	attribute	entry	category	type	description	CV_term_id	CV_term_name
2	plants	PlanteID	plantid	identifier	numeric	Plant identifier	http://purl.obolibrary.org	individual organism identifier
3	plants	Row	row	qualitative	string	Row of the invidual plant on the tabl	http://nciob.nci.nih.gov/x	Row
4	plants	Plant	plant		string	Code identifier of the individual pla	http://nciob.nci.nih.gov/x	Discrete Set Coded String Data Typ
5	plants	Treatment	treatment	factor	string	Treatment applied on plants	http://www.ebi.ac.uk/efo	environmental factor
6	plants	Genotype		qualitative	string	Genotype		
7	harvests	Lot	lot	identifier	numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
8	harvests	PlanteID			numeric	Plant identifier	http://purl.obolibrary.org	individual organism identifier
9	harvests	Truss		qualitative	string	Position on the stem of the truss	http://purl.obolibrary.org	stem node
10	harvests	HarvestDate			string	Harvest date		
11	harvests	HarvestHour			string	Harvest hour		
12	harvests	FruitAge	age	factor	string	fruit development stage	http://purl.obolibrary.org	fruit development stage
13	harvests	FruitPosition		qualitative	numeric	Poistion on the truss of the fruit	http://nciob.nci.nih.gov/x	Position Number
14	harvests	FruitDiameter		quantitative	numeric	Fruit diameter (mm)	http://nciob.nci.nih.gov/x	Diameter
15	harvests	FruitHeight		quantitative	numeric	Fruit height (mm)	http://nciob.nci.nih.gov/x	Height
16	harvests	FruitFW		quantitative	numeric	Fruit Fresh Weight(g)	http://nciob.nci.nih.gov/x	Weight
17	samples	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org	centrally registered identifier
18	samples	Lot			numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
19	samples	NbFruit			numeric	Fruit Number per sample		
20	samples	GellyFW		quantitative	numeric	Gelly Fred Weight(g) per sample		
21	samples	GellyFruit		quantitative	numeric	Gelly per Fruit (estimated g)		
22	samples	BER			string	BER		
23	compounds	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org	centrally registered identifier
24	compounds	DPA		factor	numeric	Day Per Anthesis		
25	compounds	MassBefore		quantitative	numeric	m av.extraction (g)		
26	compounds	MassMIA		quantitative	numeric	masse MIA (g)		
27	compounds	RDT		quantitative	numeric	Rdt (% MIA/DW)		
28	compounds	Starch1		quantitative	numeric	Dosage amidon (%poids/MIA)	http://purl.obolibrary.org	starch
29	compounds	Starch2		quantitative	numeric	amidon (g/gDW)	http://purl.obolibrary.org	starch
30	compounds	RHAMNOSE		quantitative	numeric	RHAMNOSE	http://purl.obolibrary.org	rhamnose

Entities

	A	B	C	D	E	F	G	H
1	rank	obtainedFrom	subset	identifier	file	description	CV_term_id	CV_term_name
2	1	0	plants	PlanteID	plants.tsv	Plant features	http://purl.obolibrary.org/obo/PO_0000003	whole plant
3	2	1	harvests	Lot	harvests.tsv	Harvest features	http://purl.obolibrary.org/obo/OBI_1110046	organ harvesting
4	3	2	samples	SampleID	samples.tsv	Samples features	http://purl.obolibrary.org/obo/PO_0009001	fruit
5	4	3	compounds	SampleID	compounds.tsv	Compound quantifications	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
6	5	3	enzymes	SampleID	enzymes.tsv	Enzyme Features	http://purl.obolibrary.org/obo/OBI_0000427	enzyme

~~Optional~~

Entity ⇔ Observational entity (e.g. samples, compounds, ...)
Attribute ⇔ Variable, Feature, ... (e.g. Plants, Fruits, Glucose, Rank, ...)

Opportunity for the Web Of Data

Resource Description Framework (RDF)

Metadata

CV Term

Attributes

	A	B	C	D	E	F	G	H
1	subset	attribute	entry	category	type	description	CV_term_id	CV_term_name
2	plants	PlanteID	plantid	identifier	numeric	Plant identifier	http://purl.obolibrary.org	individual organism identifier
3	plants	Row	row	qualitative	string	Row of the invidual plant on the tab	http://nciob.nci.nih.gov/s	Row
4	plants	Plant	plant		string	Code identifier of the individual pla	http://nciob.nci.nih.gov/s	Discrete Set Coded String Data Ty
5	plants	Treatment	treatment	factor	string	Treatment applied on plants	http://www.ebi.ac.uk/efc	environmental factor
6	plants	Genotype		qualitative	string	Genotype		
7	harvests	Lot	lot	identifier	numeric	Pool of several harvests	http://www.ebi.ac.uk/efc	sample pooling
8	harvests	PlanteID			numeric	Plant identifier	http://purl.obolibrary.org	individual organism identifier
9	harvests	Truss		qualitative	string	Position on the stem of the truss	http://purl.obolibrary.org	stem node
10	harvests	HarvestDate			string	Harvest date		
11	harvests	HarvestHour			string	Harvest hour		
12	harvests	FruitAge	age	factor	string	fruit development stage	http://purl.obolibrary.org	fruit development stage
13	harvests	FruitPosition		qualitative	numeric	Poistion on the truss of the fruit	http://nciob.nci.nih.gov/s	Position Number
14	harvests	FruitDiameter		quantitative	numeric	Fruit diameter (mm)	http://nciob.nci.nih.gov/s	Diameter
15	harvests	FruitHeight		quantitative	numeric	Fruit height (mm)	http://nciob.nci.nih.gov/s	Height
16	harvests	FruitFW		quantitative	numeric	Fruit Fresh Weight(g)	http://nciob.nci.nih.gov/s	Weight
17	samples	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org	centrally registered identifier
18	samples	Lot			numeric	Pool of several harvests	http://www.ebi.ac.uk/efc	sample pooling
19	samples	NbFruit			numeric	Fruit Number per sample		
20	samples	GellyFW		quantitative	numeric	Gelly Fred Weight(g) per sample		
21	samples	GellyFruit		quantitative	numeric	Gelly per Fruit (estimated g)		
22	samples	BER			string	BER		
23	compounds	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org	centrally registered identifier
24	compounds	DPA		factor	numeric	Day Per Anthesis		
25	compounds	MassBefore		quantitative	numeric	m av.extraction (g)		
26	compounds	MassMIA		quantitative	numeric	masse MIA (g)		
27	compounds	RDT		quantitative	numeric	Rdt (% MIA/DW)		
28	compounds	Starch1		quantitative	numeric	Dosage amidon (%poids/MIA)	http://purl.obolibrary.org	starch
29	compounds	Starch2		quantitative	numeric	amidon (g/gDW)	http://purl.obolibrary.org	starch
30	compounds	RHAMNOSE		quantitative	numeric	RHAMNOSE	http://purl.obolibrary.org	rhamnose

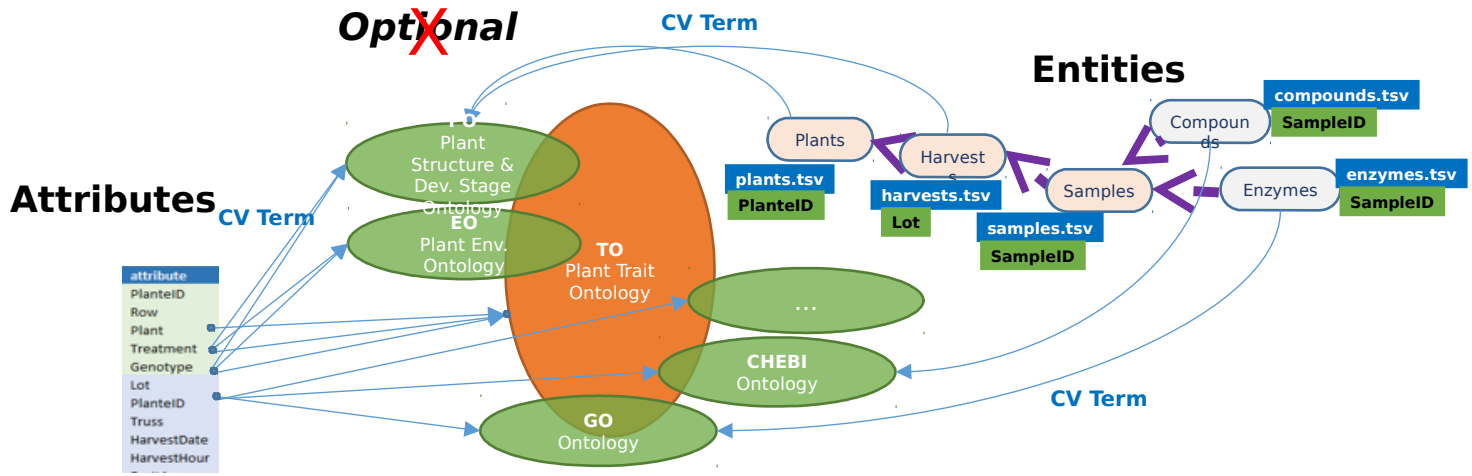
Entities

	A	B	C	D	E	F	G	H
1	rank	obtainedFrom	subset	identifier	file	description	CV_term_id	CV_term_name
2	1	0	plants	PlanteID	plants.tsv	Plant features	http://purl.obolibrary.org/obo/PO_0000003	whole plant
3	2	1	harvests	Lot	harvests.tsv	Harvest features	http://purl.obolibrary.org/obo/OBI_1110046	organ harvesting
4	3	2	samples	SampleID	samples.tsv	Samples features	http://purl.obolibrary.org/obo/PO_0009001	fruit
5	4	3	compounds	SampleID	compounds.tsv	Compound quantifications	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
6	5	3	enzymes	SampleID	enzymes.tsv	Enzyme Features	http://purl.obolibrary.org/obo/OBI_0000427	enzyme

Reference ontologies

a conceptualization associated with a set of facts (domain-dependent)

AgroPortal LIRMM http://agroportal.lirmm.fr/ontologies

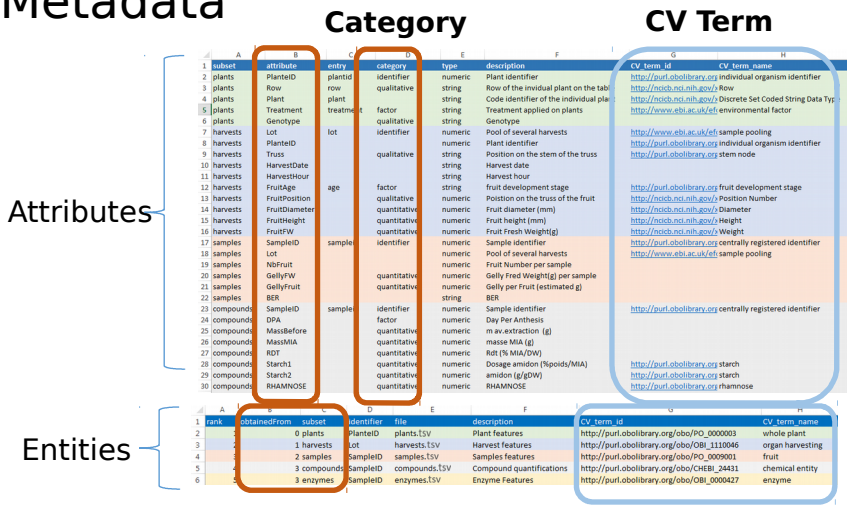


Entity ⇔ Observational entity (e.g. samples, compounds, ...)
Attribute ⇔ Variable, Feature, ... (e.g. Plants, Fruits, Glucose, Rank, ...)

FAIR principles, a new opportunity to improve the data lifecycle

Opportunity for the Web Of Data

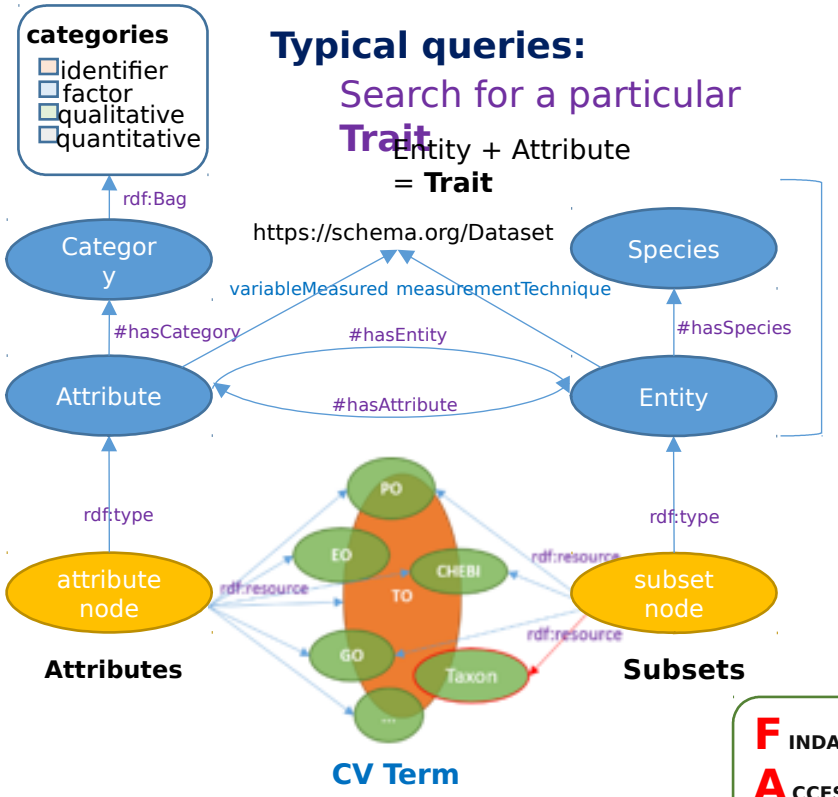
Metadata



Resource Description Framework (RDF)

Application ontologies

a conceptual model or ontologies within a knowledge base.



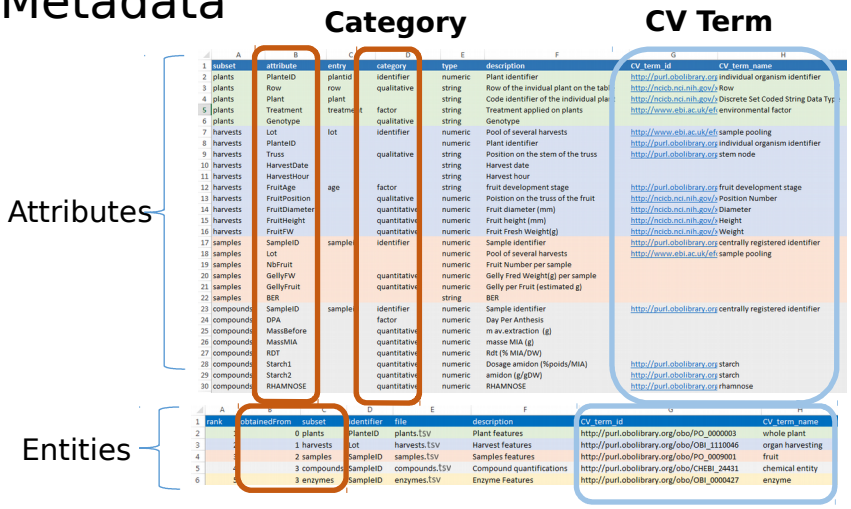
100 %

F INDABLE
A CCESSIBLE
I NTEROPERABLE
R E-USABLE

FAIR principles, a new opportunity to improve the data lifecycle

Opportunity for the Web Of Data

Metadata



Other Metadata

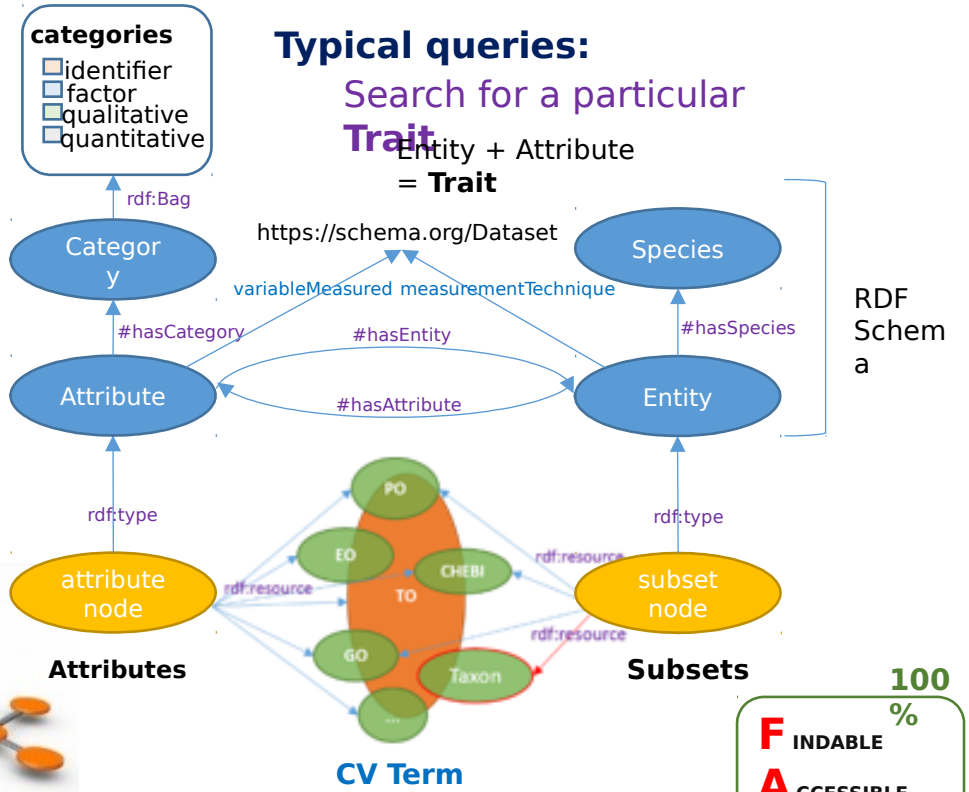
images,
geolocation,
meteorology,
other data sets,
...

Linked Data

Resource Description Framework (RDF)

Application ontologies

a conceptual model or ontologies within a knowledge base.



100 %

F INDABLE
A CCESSIBLE
I NTEROPERABLE
R E-USABLE

FAIR principles

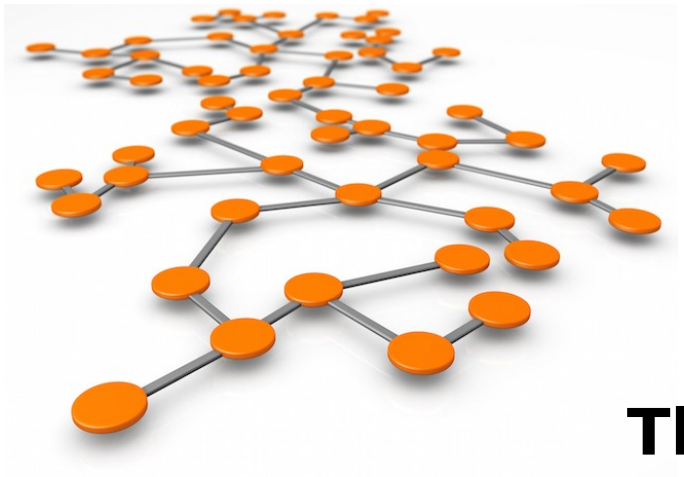
- can **improve** the **data lifecycle** by relying on technologies,
- but **only** if they **serve** the concerned **communities by meeting their expectations.**

FAIR principles

- can **improve** the **data lifecycle** by relying on technologies,

- but **only** if they **serve** the concerned **communities by meeting their expectations.**

The role of a data authority is to translate principles into standards in agreement with the concerned communities



Thank you for your attention

FAIR PROSODie

INRA SCIENCE & IMPACT