ODAM Framework
Open Data for Access and Mining

EDTMS

How to best manage your data to make the most of it for your research ?

How to ensure that open data works for research ?

Preliminary question :

Data Capture ?

The main focus is on building framework and solution to store data

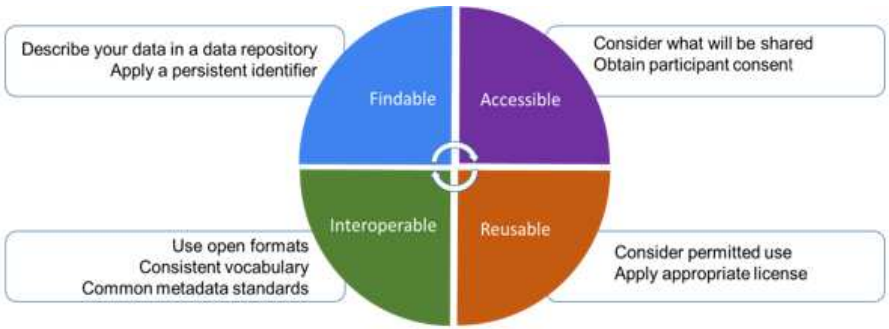
Daniel Jacob

INRA UMR 1332 BFP – Metabolism Group

Bordeaux Metabolomics Facility

INRA
SCIENCE & IMPACT

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

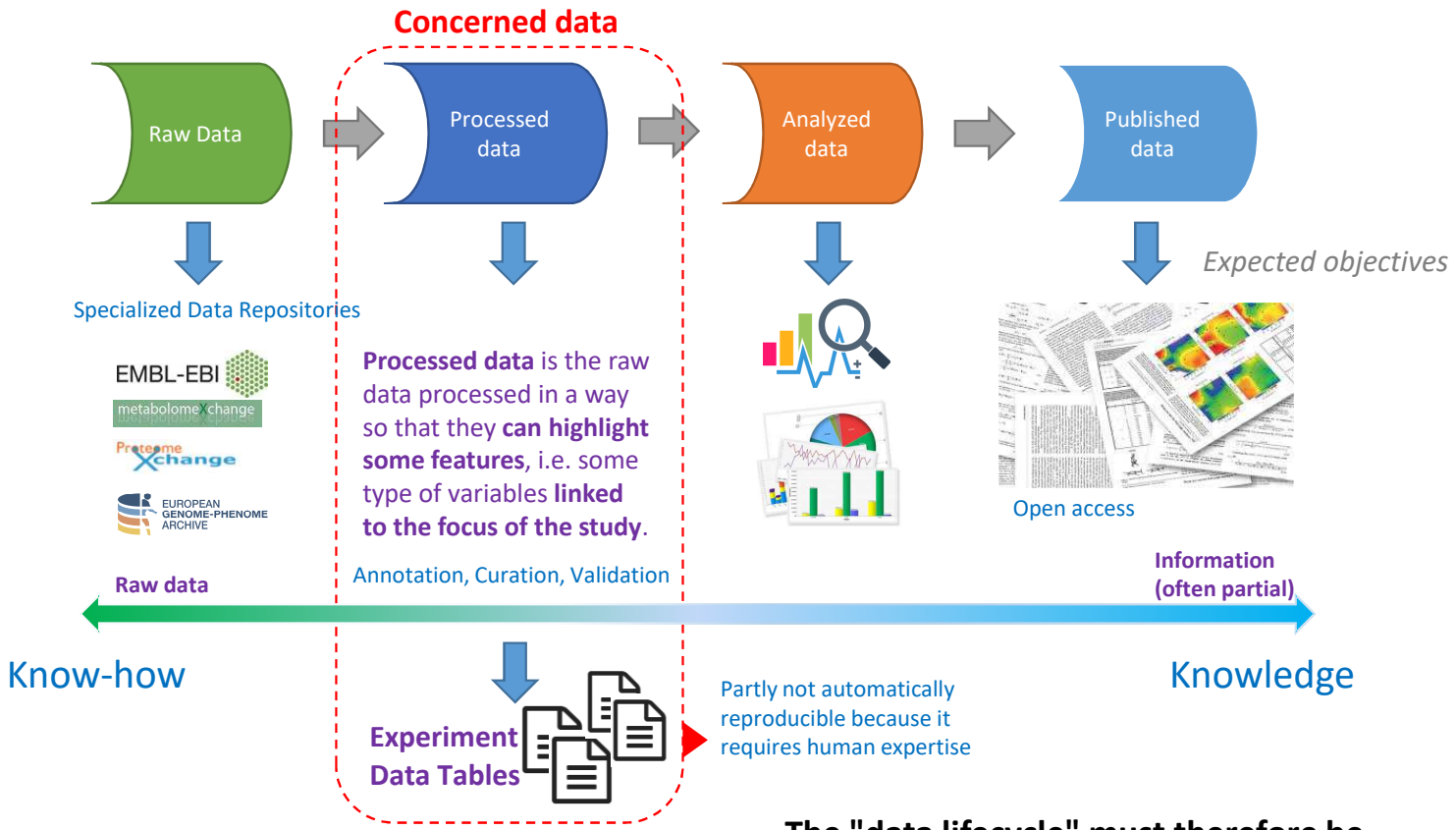
The implementation of FAIR principles is a process that must be thought of in a progressive and community-oriented way.

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

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

Data flow

Studies with a design of experiment (DoE)

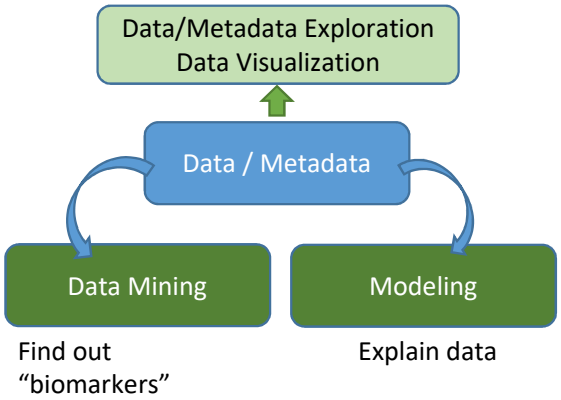


⇨ Data capture ...
as far upstream as possible

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

Data flow

Data processing



Know-how

During a research project

DATA

Project

Data exploration ⇨ Descriptive statistics

First glimpse of the data that can show trends.

Allow the data to be well characterized, which is necessary to then choose how to analyze them.

Data mining / Modeling

- ⇨ Repetition of multiple scenarios on different subsets of data
- ⇨ Selection subsets of data

⇨ **Implying lots of data manipulation**

- ⇨ **data capturing, data formatting, data linking, data import / data export**
- ⇨ **Linking both metadata and data for data mining**

Knowledge

Data flow

Data management

HORIZON 2020

European Commission

Publishing policies

Before the project begins

DMP OPIDoR

During a research project

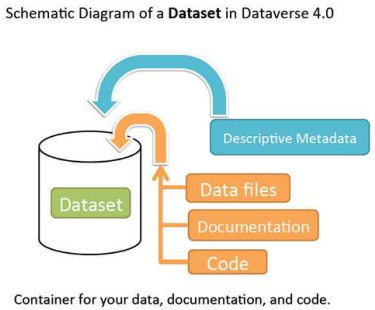
DATA

After the project is completed

The Dataverse Project

Data

Future



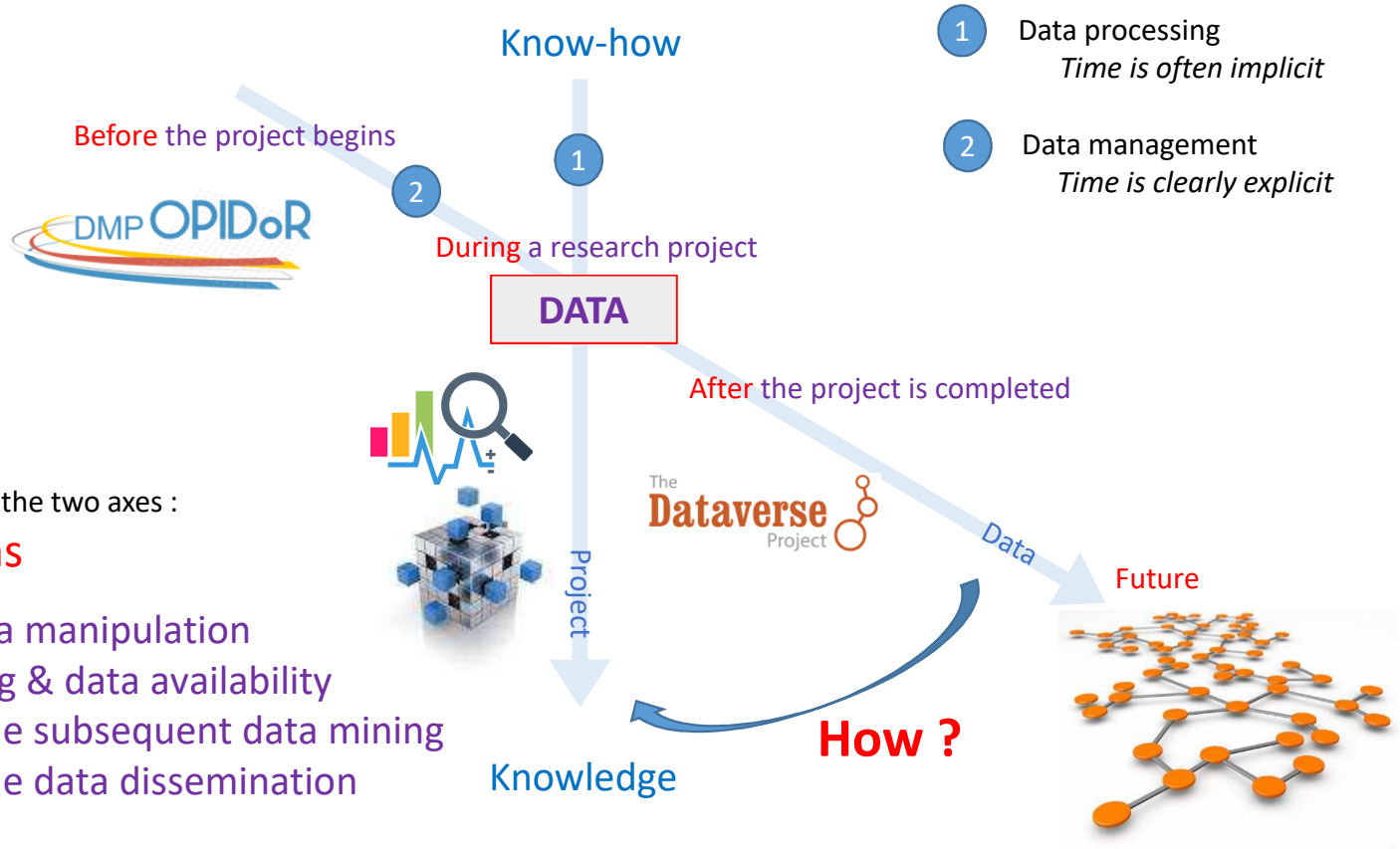
A **data management plan or DMP** is a formal document that outlines how data are to be handled both during a research project, and after the project is completed.

The goal of a data management plan is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins

this ensures that data are well-managed in the present, and prepared for preservation in the future.

- ⇨ **Description by metadata (Ontology / Controlled Vocabulary)**
- ⇨ **Data capturing, data formatting, data linking**
- ⇨ **Implying data archeology after several months / years**

Data flow



Make consistent the two axes :

Motivations

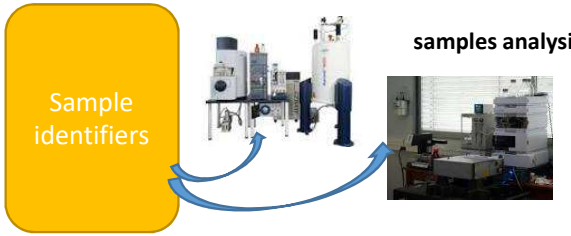
- ⇨ Reduce data manipulation
- ⇨ Data sharing & data availability
- ⇨ Facilitate the subsequent data mining
- ⇨ Facilitate the data dissemination

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

Use-Case

Experiment Data Tables

Experiment Design



plants.tsv

Genotype	PlanteID	Plant	Row	Treatment
Money Maker_WT	26	A26	A	Control
Money Maker_WT	140	C2	C	Control
Money Maker_WT	222	D15	D	Control
Money Maker_WT	295	E19	E	Control
Money Maker_WT	319	F24	F	Control

harvests.tsv

PlanteID	Lot	Truss	FruitAge	HarvestDate	HarvestHour	FruitPositior	FruitFW	FruitDiamet	FruitHeight
18	5	Truss_5	00.08DPA	40379	0.51875	4	1.1	13.29	13.17
155	5	Truss_5	00.08DPA	40379	0.51875	4	1.01	12.38	12.29
164	5	Truss_5	00.08DPA	40379	0.51875	4	1.02	12.28	12.44
221	5	Truss_5	00.08DPA	40379	0.51875	2	1.02	12.8	12.28
226	5	Truss_5	00.08DPA	40379	0.51875	3	0.81		
322	5	Truss_5	00.08DPA	40379	0.51875	2	0.79		
343	5	Truss_5	00.08DPA	40379	0.51875	4	0.69		
361	5	Truss_5	00.08DPA	40379	0.51875	5	1.06		
369	5	Truss_5	00.08DPA	40379	0.51875	4	0.73		
372	5	Truss_5	00.08DPA	40379	0.51875	5	1.21		

samples.tsv

Lot	SampleID	NbFruit	GellyFW	GellyFruit	BER
1	1	10	2.83	0.283	FALSE
1	2	10	2.83	0.283	FALSE
2	3	8	4.05	0.50625	FALSE
3	5	10	3.3	0.33	FALSE
3	6	10	3.3	0.33	FALSE

compounds.tsv

SampleID	DPA	MassBefore	MassMIA	RDT	Starch1	Starch2	RHAMNOSE	ARABINOSE	XYLOSE	MANNOSE	GALACTOSE	GLUCOSE	OsesN	PoidsAU	PoidsTotal	Pf
363	15	0.192	0.1002	52.19	36.55	0.19074531	0.89	2.17	1.73	1.2	5.99	11.31	23.29	7.17	30.46	
369	15	0.107	0.0511	47.76	32.43	0.15487598	0.74	2.19	1.68	1.3	6.03	8.51	20.45	7.35	27.8	
373	15	0.166	0.0771	46.45	32.05	0.14885874	0.93	2.44	2.56	2.95	9.3	18.18	36.36	10.88	47.24	
375	15	0.14	0.0684	48.86	34.89	0.17046257	0.71	1.89	1.98	2.04	7.33	10.47	24.42	8.79	33.21	
429	28	0.104	0.0431	41.44	19											
431	28	0.169	0.0841	49.76	34											
433	15	0.187	0.0916	48.98	3											
435	15	0.179	0.092	51.4	35											
445	26.7	0.156	0.0612	39.23	18											
451	28	0.09	0.035	38.89	18											
457	28	0.148	0.0707	47.77	31											
463	26.7	0.106	0.0415	39.15	19											
531	28	0.152	0.0555	36.51	19											
533	28	0.107	0.0419	39.16	24											
1025	53	0.126	0.0344	27.3	1											
1027	53	0.092	0.0252	27.39												
1057	56	0.081	0.0182	22.47												
1067	53	0.089	0.022	24.72	0											
1069	53	0.134	0.0363	27.09	0											
1103	26.7	0.098	0.0368	37.55	13											
1131	56	0.116	0.0257	22.16	0											
1163	53	0.108	0.0248	22.96	0											

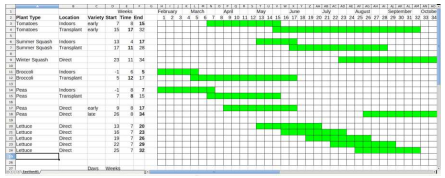
enzymes.tsv

SampleID	PGM	F16BP_Cyt	PyrK	CitS	PPI	AcoS	PFK	FruS	F16BP_Stron	GluS	ISODH_NAD	EnoS	ISODH_NADI	PEPC
1	NA	8.97	1599.53	64.53	2767.89	1172.28	192.05	876.13	523.57	722.19	508.99	NA	NA	93
3	6844.78	85.02	1839.39	NA	3373.46	2014.02	263.33	984.08	634.14	622.22	975.77	514.4	133.6	11
5	5336.12	14.58	1399.41	65.84	2642.21	980.02	184.31	743.8	445.98	518.01	555.54	588.57	167.86	99
7	5376.43	23.81	1333.66	105.67	2663.67	1349.45	147.18	807.87	501.92	915.19	572.33	676.61	87.43	27
9	5230	25.21	1271.13	110.25	2575.94	865.05	177.67	792.12	459.77	510.8	421.23	504.23	231.52	81
11	6147.04	24.19	754.24	68.66	2272.79	1005.74	144.17	641.46	467.71	598.68	416.04	530.9	59.01	8
109	5416.24	35.57	1833.13	168.54	2779.48	869.08	197.73	971.88	494.3	614.79	660.68	645.49	221.32	68
111	8054.61	33.14	2128.92	64.59	3128.5	1830.77	270.3	1008.07	778.47	1271.77	965.12	530.51	68.32	9
113	3997.96	0.7	3053.11	44.2	2330.09	719.39	250.87	1223.19	360.15	526.31	437.56	548.96	194.09	84
115	5459.74	23.73	1944.78	59.71	3338.61	1168.66	246.51	1039.7	526.42	606.25	607.02	557.28	163.25	1
117	7509.32	1.71	1593.34	8.34	2597.87	1035.6	260.67	1019.02	663.14	986.63	632.42	579.25	NA	90
119	6018.59	21.54	1624.31	50.1	2755.34	1245.73	243.07	915.78	530.86	832.47	659.72	518.63	91.02	87
361	3967.66	12.54	786.54	88.85	1828.44	363.23	131.41	525.24	268.55	337.81	210.84	581.39	118.38	61
363	4753.73	17.77	728.5	NA	1869.46	450.13	171.89	593.3	357.74	220.39	247.36	587.58	20.21	86
365	3836.19	4.16	603.74	31.55	1836.44	330.26	131.87	499.38	208.85	249.95	190.41	545.06	58.85	57
367	4.11	4.87	507.67	7.39	1371.13	573.06	128.65	390.78	283.25	304.36	301.7	NA	168.46	83
369	3798.59	15.14	703.9	82.78	1473.25	286.02	120.11	438.09	260.44	240.08	204.19	500.82	88.22	
371	5610.69	14.69	1092.42	99.09	2384.35	799.55	182.18	812.09	425.56	583.97	366.29	639.19	45.01	127
373	4220.68	31.7	592.57	14.81	1654.27	602.62	150.23	480.9	283.92	354.94	256.53	766.52	NA	6
375	3342.92	13.68	557.38	1.32	1612.42	359.41	119.45	407.36	279.22	290.21	202.93	600.66	41.36	9
377	3100.01	21.21	467.95	60.43	1095.82	369.89	93.32	206.48	170.71	196.26	146.66	660.62	124.38	58

Research question ⇨ Project ⇨ Experiment ⇨ Experimental set-up

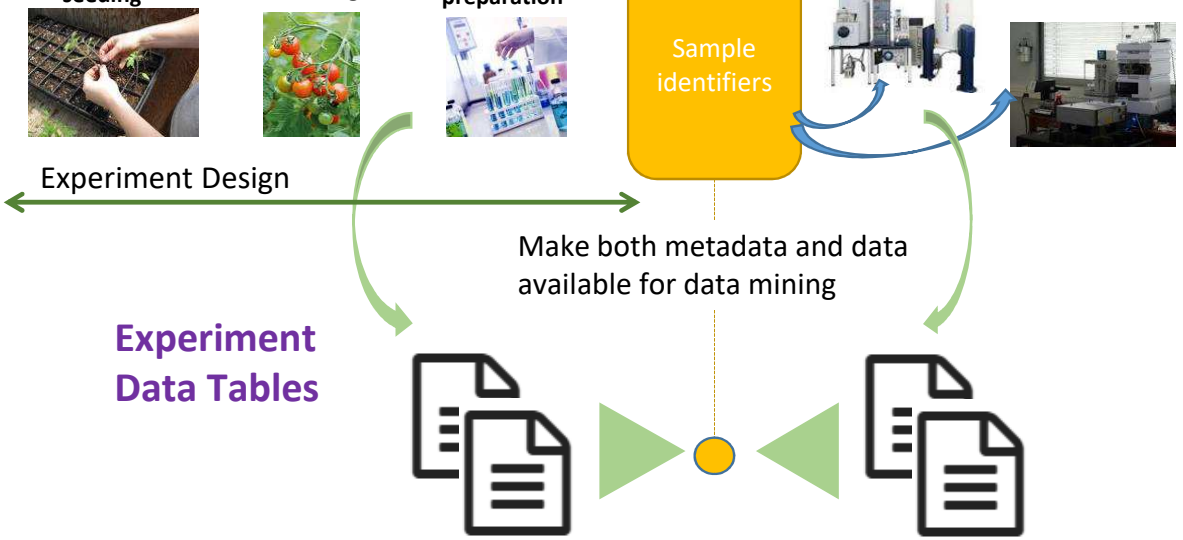
Before the project begins

project building
then experiment
design



Experiment's Mapping

During a research project

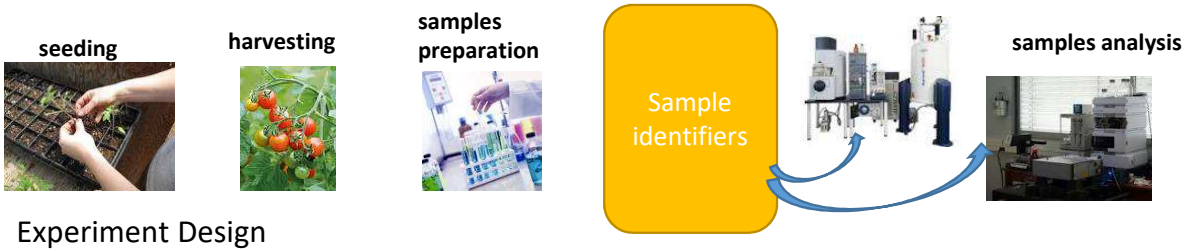


Several operators, technics, data types, SOPs, …

Several partners

Each time we plan to **share data** coming from a common experimental design,
the classical challenges for fast using data by every partner are **data storage and data access**

The experimental context: recurrent needs / wishes



Experiment Design

1. data sharing & data availability
2. identifiers centrally managed
3. facilitate the subsequent publication of data
4. avoid the implementation of a RDBDMS

ODAM Framework
Open Data for Access and Mining

Purposes

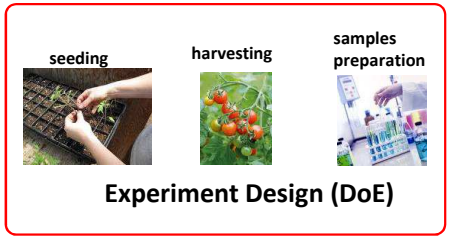
Objectives

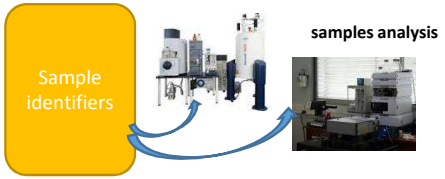
- make research **data locally or broadly accessible all along the project**
- allow any (data) scientists to be able to **explore the dataset** and then **extract a subpart or the totality of the data** according to their needs
- allow data to be selected then, **downloadable by web API**
- allow data and analysis to be **visualized online**

Guideline keywords

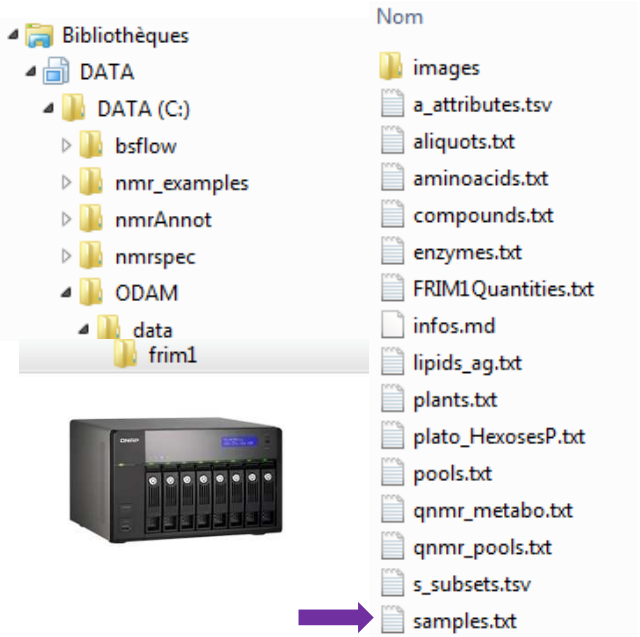
- simplicity, flexibility, efficiency

Use-Case





ODAM Framework
Open Data for Access and Mining

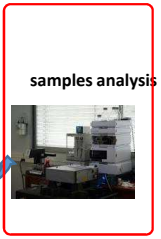


A	B	C	D	E	F	G	H	I	J
SampleID	Treatment	DevStage	FruitAge	FruitPosition	FruitDiamete	FruitHeight	FruitFW	Rank	Truss
115	Control	FF.01	07DPA	3	11.95	10.42	0.81	A	T7
121	Control	FF.03	22DPA	3	36.13	31.77	21.43	A	T6
164	Control	FR.01	42DPA	2	51.09	46.85	64.05	A	T5
353	Control	FR.04	55DPA	5	48.28	43.35	66.64	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

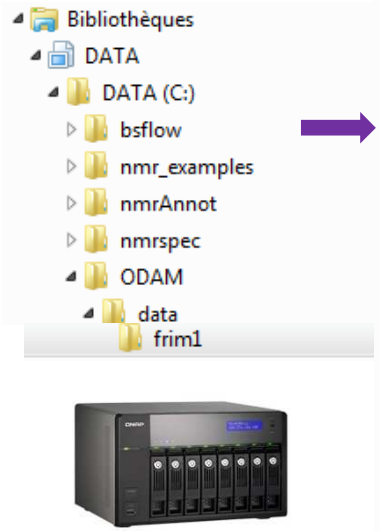
Use-Case

Experiment Design (DoE)

Sample identifiers



ODAM Framework
Open Data for Access and Mining



A	B
SampleID	AliquotID
1	1
1	2
2	3
2	4
3	5
3	6
4	7
4	8
5	9
5	10
6	11
6	12
31	109
31	110
32	111
32	112
33	113
33	114
34	115
34	116
35	117
35	118
36	119
36	120
97	361
97	362

aliquots.txt

Use-Case


seeding


harvesting


samples preparation

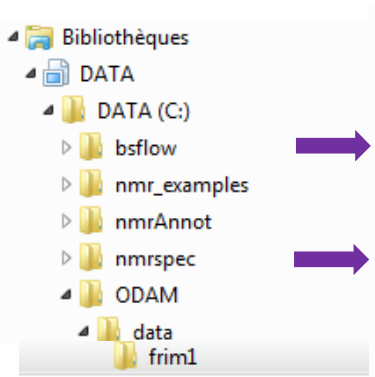
Sample identifiers

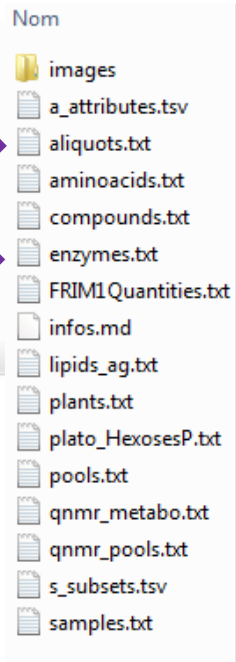

samples analysis

Experiment Design (DoE)

ODAM Framework
Open Data for Access and Mining


EDTMS





A	B
SampleID	AliquotID
1	1
1	2
2	3
2	4
3	5
3	6
4	7
4	8
5	9
5	10
6	11
6	12
31	109
31	110
32	111
32	112
33	113
33	114
34	115
34	116
35	117
35	118
36	119
36	120
97	361
97	362

aliquots.txt

A	B	C	D	E	F	G	H	I	J	K	L
AliquotID	PGM	F16BP_Cyt	PyrK	CitS	PPI	AcoS	PFK	FruS	F16BP_Stron	GluS	ISODH_NAD
1	NA	8.97	1599.53	64.53	2767.89	1172.28	192.05	876.13	523.57	722.19	508.99
3	6844.78	85.02	1839.39	NA	3373.46	2014.02	263.33	984.08	634.14	622.22	975.77
5	5336.12	14.58	1399.41	65.84	2642.21	980.02	184.31	743.8	445.98	518.01	555.54
7	5376.43	23.81	1333.66	105.67	2663.67	1349.45	147.18	807.87	501.92	915.19	572.33
9	5230	25.21	1271.13	110.25	2575.94	865.05	177.67	792.12	459.77	510.8	421.23
11	6147.04	24.19	754.24	68.66	2272.79	1005.74	144.17	641.46	467.71	598.68	416.04
109	5416.24	35.57	1833.13	168.54	2779.48	869.08	197.73	971.88	494.3	614.79	660.68
111	8054.61	33.14	2128.92	64.59	3128.5	1830.77	270.3	1008.07	778.47	1271.77	965.12
113	3997.96	0.7	3053.11	44.2	2330.09	719.39	250.87	1223.19	360.15	526.31	437.56
115	5459.74	23.73	1944.78	59.71	3338.61	1168.66	246.51	1039.7	526.42	606.25	607.02
117	7509.32	1.71	1593.34	8.34	2597.87	1035.6	260.67	1019.02	663.14	986.63	632.42
119	6018.59	21.54	1624.31	50.1	2755.34	1245.73	243.07	915.78	530.86	832.47	659.72
361	3967.66	12.54	786.54	88.85	1828.44	363.23	131.41	525.24	268.55	337.81	210.84
363	4753.73	17.77	728.5	NA	1869.46	450.13	171.89	593.3	357.74	220.39	247.36
365	3836.19	4.16	603.74	31.55	1836.44	330.26	131.87	499.38	208.85	249.95	190.41
367	4.11	4.87	507.67	7.39	1371.13	573.06	128.65	390.78	283.25	304.36	301.7
369	3798.59	15.14	703.9	82.78	1473.25	286.02	120.11	438.09	260.44	240.08	204.19
371	5610.69	14.69	1092.42	99.09	2384.35	799.55	182.18	812.09	425.56	583.97	366.29
373	4220.68	31.7	592.57	14.81	1654.27	602.62	150.23	480.9	283.92	354.94	256.53
375	3342.92	13.68	557.38	1.32	1612.42	359.41	119.45	407.36	279.22	290.21	202.93
377	3100.01	21.21	467.95	60.43	1095.82	369.89	93.32	206.48	170.71	196.26	146.66
379	NA	NA	584.29	26.11	1787.02	668.52	158.95	438.82	262.62	334.18	291.38
381	3483.95	8.98	771.26	13.32	1514.18	497.3	115.14	337.25	234.69	271.34	216.68
384	4011.16	4.85	590.28	13.75	1245.94	378.87	109.77	344.62	193.74	165.72	196.86
385	1767.14	11.97	351.52	43.95	881.43	174.81	85.89	91.79	127.57	82.19	135.55
387	1907.36	10.54	268.66	36.35	964.63	135.77	70.58	137.41	108.43	151.45	110.26
389	1843.23	10.37	214.02	22.79	837.37	82.88	85.74	141.82	124.58	61.97	120.31
391	2322.24	15.06	349.62	52.12	1014.67	230.62	80.48	185.05	187.63	56.33	151.72
393	3826.41	18.98	466.4	103.64	1485.31	212.96	140.72	304.75	184.27	181.77	169.72
395	2483.08	18.13	556.97	63.22	1154	286.7	103.31	193.56	127.9	138.77	172.3
397	4924.26	17.89	1327.67	50.79	2225.01	657.4	160.33	638.3	398.38	458.04	495.24

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).

Promote good practices

Provide services

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

Handling of all these tasks related to data management

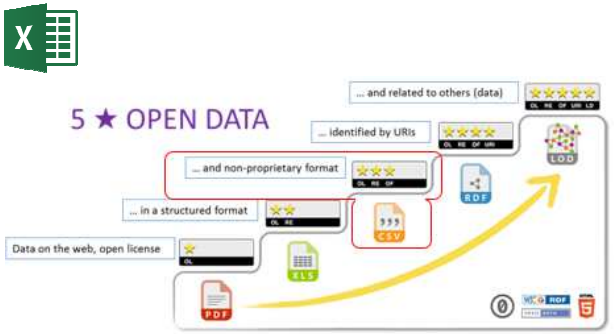
Promote good practices

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** ».

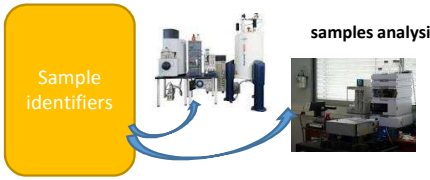
Promote non-proprietary format like CSV or TSV

Well organized data

- Each variable forms a column
- Each observation forms a line
- Each type of "unit observational" forms a table

≠ notepad !

Promote 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
		R.04	55DPA	5	48.28	43.35	66.64	A	T5
		R.04	55DPA	3	49.84	44.93	66.98	A	T5
		R.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			3	13.44	12.39	1.14	A	T7
536	Control			NA	59.4	49.05	87.28	A	TA
544	Control			NA	57.31	47.69	92.86	A	TA
158	Control	FF.04	35DPA				92.86	A	T5
109	Control	FF.03	22DPA				38.73	A	T5
134	Control	FF.02	15DPA				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		
520	Control	FF.03	30DPA	NA	48.86	41.52	52.94		
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

Identifier

Factors

Quantitatives

Qualitatives

Data

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, tissues, bacteria, …)

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.

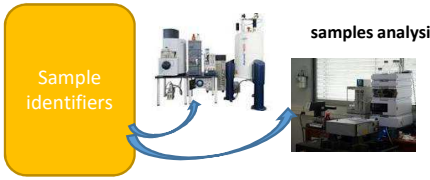
Well organized data

- Each variable forms a column
- Each observation forms a line
- Each type of "unit observational" forms a table

Promote non-proprietary format like CSV or TSV

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019

Promote 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]	FR.04	55DPA	5	48.28	43.35	66.64	A	T5
[illegible]	[illegible]	FR.04	55DPA	3	49.84	44.93	66.98	A	T5
[illegible]	[illegible]	FR.02	47DPA	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
512	Control	FF.03	21DPA	NA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
117	Control	[illegible]	[illegible]	3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
536	Control	[illegible]	[illegible]	NA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
544	Control	[illegible]	[illegible]	NA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
158	Control	FF.04	35DPA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
109	Control	FF.03	22DPA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
134	Control	FF.02	15DPA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
31	Control	FF.01	08DPA	4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
179	Control	FF.03	28DPA	3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
383	Control	FF.04	34DPA	5	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
425	Control	FR.04	55DPA	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
520	Control	FF.03	30DPA	NA	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
419	Control	FR.03	50DPA	2	[illegible]	[illegible]	[illegible]	[illegible]	[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

Qu

Data

Description of the different columns within data files

Shortname	Description	Unit	
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

categories

- identifier
- factor
- qualitative
- quantitative

Metadata

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

Promote non-proprietary format like CSV or TSV

Promote good practices

Minimal but relevant Metadata

Metadata Files

- Our approach, by **adjoining some minimal but relevant metadata**, gives access to the data themselves **with the possibility to explore and mine them**.

For that, 2 definition files for associate metadata are required

- **s_subsets.tsv**: a definition file allowing each data subset file to be associated with a concept

- **a_attributes.tsv**: a definition file allowing each attribute (concept/variable) to be annotated with some minimal but relevant metadata

ODAM Framework
Open Data for Access and Mining

EDTMS

Promote good practices

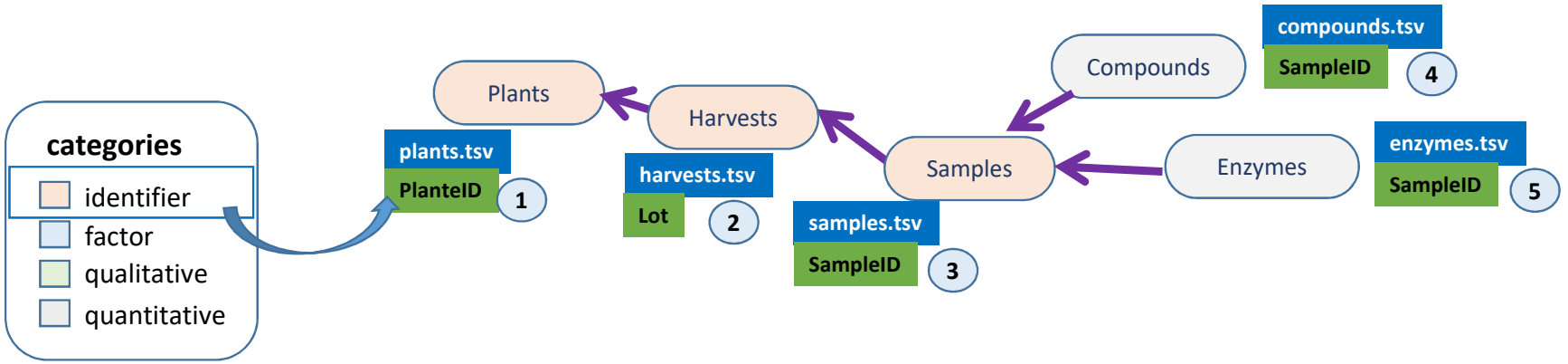
Minimal but relevant Metadata

Metadata Files

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.tsv	Plant features	http://purl.obolibrary.org/obo/PO_0000003	whole plant
3	2	1	harvests	Lot	harvests.tsv	Harvest features	http://purl.obolibr	organ harvesting
4	3	2	samples	SampleID	samples.tsv	Samples features	http://purl.obolibr	fruit
5	4	3	compounds	SampleID	compounds.tsv	Compound quantifications	http://purl.obolibr	chemical entity
6	5	3	enzymes	SampleID	enzymes.tsv	Enzyme Features	http://purl.obolibrary.org/obo/OBI_0000427	enzyme

Optional



→ "Is Obtained From"

ODAM Framework
Open Data for Access and Mining

EDTMS

Promote good practices

Minimal but relevant Metadata

Metadata Files

s_subsets.tsv

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

rank	obtainedFrom	subset	identifier	file	description	CV_term_id	CV_term_name
1	0	plants	PlanteID	plants.csv	Plant features	http://purl.obolibrary.org/obo/PO_0000003	whole plant
2	1	harvests	Lot	harvests.csv	Harvest features	http://purl.obolibrary.org/obo/PO_1110046	organ harvesting
3	2	samples	SampleID	samples.csv	Samples features	http://purl.obolibrary.org/obo/PO_0009001	fruit
4	3	compounds	SampleID	compounds.csv	Compound quantifications	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
5	4	enzymes	SampleID	enzymes.csv	Enzyme Features	http://purl.obolibrary.org/obo/OBI_0000427	enzyme

Optional

...

Entity ⇔ unit observational (e.g. samples, compounds, …)

Metadata file allowing each attribute (variable) to be annotated

a_attributes.tsv

subset	attribute	entry	category	type	description	CV_term_id	CV_term_name
plants	PlanteID	plantid	identifier	numeric	Plant identifier	http://purl.obolibrary.org/	individual organism identifier
plants	Row	row	qualitative	string	Row of the invidual plant on the table	http://ncicb.nci.nih.gov/x	Row
plants	Plant	plant		string	Code identifier of the individual plant	http://ncicb.nci.nih.gov/x	Discrete Set Coded String Data Type
plants	Treatment	treatment	factor	string	Treatment applied on plants	http://www.ebi.ac.uk/efo	environmental factor
plants	Genotype		qualitative	string	Genotype		
harvests	Lot	lot	identifier	numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
harvests	PlanteID			numeric	Plant identifier	http://purl.obolibrary.org/	individual organism identifier
harvests	Truss		qualitative	string	Position on the stem of the truss	http://purl.obolibrary.org/	stem node
harvests	HarvestDate			string	Harvest date		
harvests	HarvestHour			string	Harvest hour		
harvests	FruitAge	age	factor	string	fruit development stage	http://purl.obolibrary.org/	fruit development stage
harvests	FruitPosition		qualitative	numeric	Poistion on the truss of the fruit	http://ncicb.nci.ni	
harvests	FruitDiameter		quantitative	numeric	Fruit diameter (mm)	http://ncicb.nci.ni	
harvests	FruitHeight		quantitative	numeric	Fruit height (mm)	http://ncicb.nci.ni	
harvests	FruitFW		quantitative	numeric	Fruit Fresh Weight(g)	http://ncicb.nci.nih.gov/x	Weight
samples	SampleID	sampleid	identifier	numeric	Sample identifier	http://purl.obolibrary.org/	centrally registered identifier
samples	Lot			numeric	Pool of several harvests	http://www.ebi.ac.uk/efo	sample pooling
samples	NbFruit			numeric	Fruit Number per sample		
samples	GellyFW		quantitative	numeric	Gelly Fred Weight(g) p		
samples	GellyFruit		quantitative	numeric	Gelly per Fruit (estima		
samples	BER			string	BER		
compounds	SampleID	sampleid	identifier	numeric	Sample identifier		
compounds	DPA		factor	numeric	Day Per Anthesis		
compounds	MassBefore		quantitative	numeric	m av.extraction (g)		
compounds	MassMIA		quantitative	numeric	masse MIA (g)		
compounds	RDT		quantitative	numeric	Rdt (% MIA/DW)		
compounds	Starch1		quantitative	numeric	Dosage amidon (%poids/MIA)	http://purl.obolibrary.org/	starch
compounds	Starch2		quantitative	numeric	amidon (g/gDW)	http://purl.obolibrary.org/	starch
compounds	RHAMNOSE		quantitative	numeric	RHAMNOSE	http://purl.obolibrary.org/	rhamnose

Optional

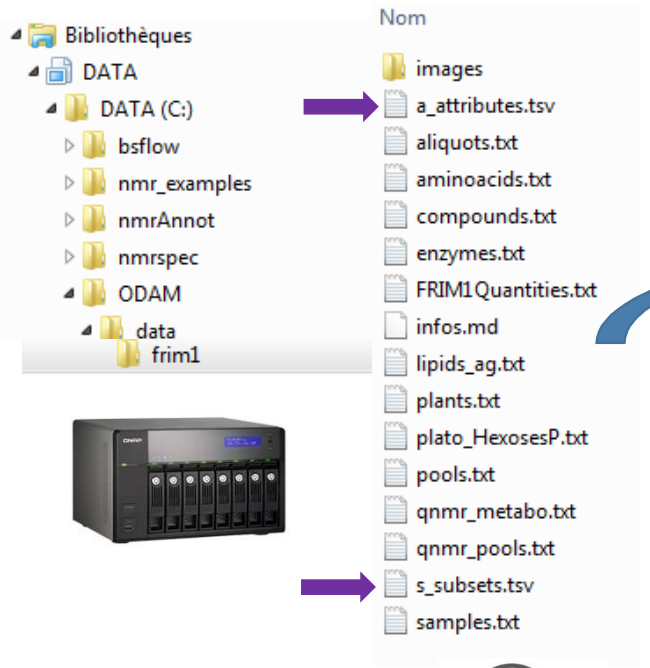
...

categories
- identifier
- factor
- qualitative
- quantitative

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

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

Data Emancipation



ODAM

EDTMS

F
A
I NTEROPERABLE
R

With the help of the two metadata files

⇒ Data emancipation regarding Tools

Data ⇔API ⇔ Tools

Data — API — **Tools**

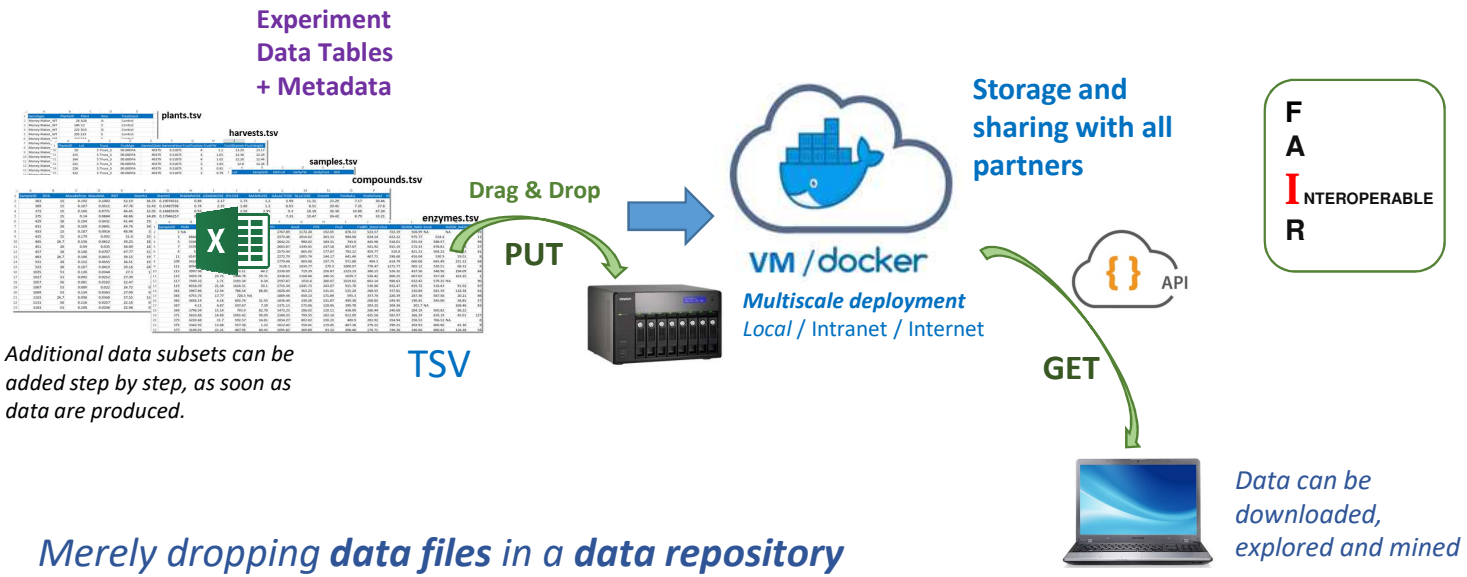
Multiscale deployment
Local / Intranet / Internet

- ODAM framework allows experimental data tables to be widely accessible and fully reusable including through a scripting language such as R, and this with minimal effort on the part of the data provider.
- The approach consists in building a web-based data network, based on appropriate technologies (Web API), and using standard data formats (TSV, JSON).
- Web applications, each with a clearly defined objective, then operate this network.
- A data can therefore be used for several applications and vice versa. The data management system becomes completely independent of its operation.
- The data is thus “decompartmentalized", a sine qua non condition for the Web of Data

ODAM Framework Overview

Promote good practices → **Provide services**

ODAM
Open Data for Access and Mining



*Merely dropping **data files** in a **data repository** (e.g. a local disk, NAS or distant storage space) should allow users to **access them by web API***

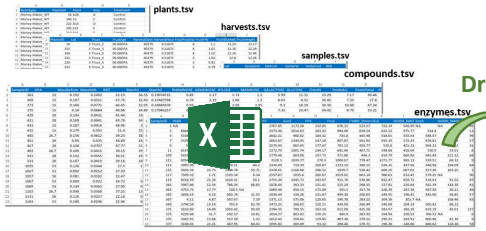
No database schema, no programming code are required

ODAM Framework Overview

Promote good practices → **Provide services**

ODAM
Open Data for Access and Mining

Experiment Data Tables + Metadata



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

Drag & Drop

PUT

Multiscale deployment
Local / Intranet / Internet

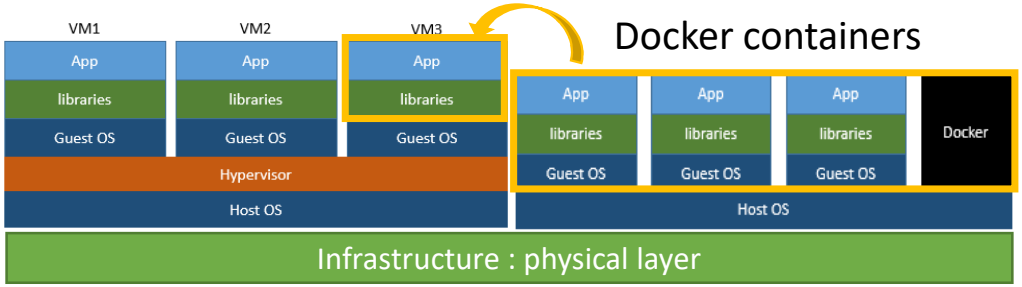
Storage and sharing with all partners

GET

Data can be downloaded, explored and mined

F
A
I NTEROPERABLE
R

Virtual Machines



Docker containers

Infrastructure : physical layer

Infrastructure : PC Desktop, Laptop, Server, Data centre, Cloud
Host OS : MS Windows 10, Mac OS, Unix / Linux, …
Hypervisor : Virtual Box, VMware, …

Multiscale deployment
Local / Intranet / Internet

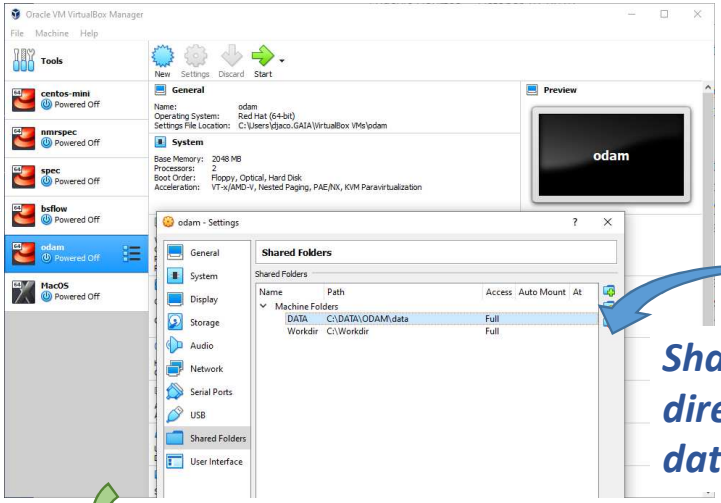
Daniel Jacob – INRA UMR 1332 BFP – Oct 2019

ODAM
Virtual Disk Image

1

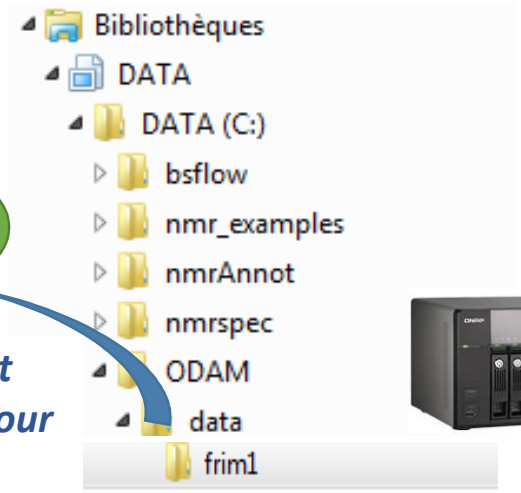
Multiscale deployment
Local / Intranet / Internet

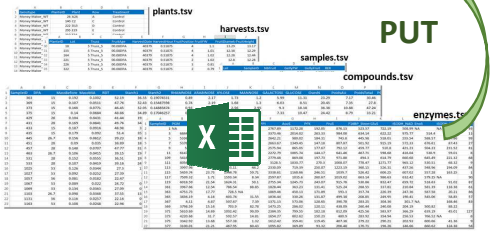
Infrastructure : PC Desktop, Laptop
Host OS : MS Windows 10, Mac OS
Hypervisor : Virtual Box



Share the root directory of your datasets

2





PUT

TSV

API

GET

Data can be downloaded, explored and mined

3

See the ODAM installation Guide

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

https://bio.tools/ODAM

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019

Application
Programming
Interface

Make both metadata and data available for data mining

ODAM

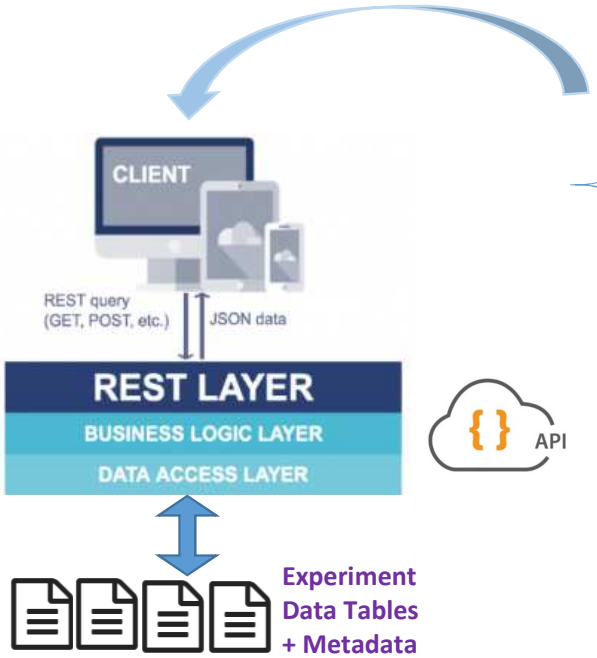
EDTMS

Open Data for Access and Mining 1.0.0 OAS3

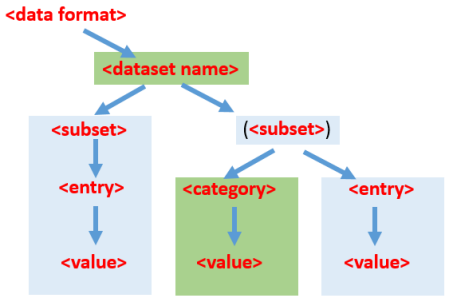
https://pmb-bordeaux.fr/odamsw/

REST Services: hierarchical tree of resource naming (URL)

With the help of the two metadata files



GET	`/getdata/infos/{dataset}`	get information
GET	`/getdata/json/{dataset}/check`	test checks
GET	`/getdata/{format}/{dataset}`	Get the subset list of a dataset
GET	`/getdata/{format}/{dataset}/metadata`	Get all attribute metadata
GET	`/getdata/{format}/{dataset}/({subset})?limit={limit}`	Get all values of a merged data subsets
GET	`/getdata/{format}/{dataset}/({subset})/entry`	Get the entry list of a merged data subsets
GET	`/getdata/{format}/{dataset}/({subset})/{entry}/{value}?limit={limit}`	Get all values of a merged data subsets for a specific value of an (WS)entry
GET	`/getdata/{format}/{dataset}/({subset})/{category}`	Get the variable list within the specified category of a merged data subsets



Data can be explored and mined based on some **minimal but relevant metadata**

API

Open Data for Access and Mining

https://pmb-bordeaux.fr/odamsw/

ODAM

EDTMS

Request URL

```
http://pmb-bordeaux.fr/getdata/tsv/friml/(activome,qNMR_metabo)/sample/365?limit=10
```

Identifiers

PlantID	Rank	PlantNum	Treatment	SampleID	Truss	DevStage	FruitAge	AliquotID	PGM	F16BP_Cyt	PyrK	CitS	PPI	AcoS	PoolID	glucose	saccharose	fructose	galactose
A1	A	1	Control	115	T7	FF.01	07DPA	397	4974.26	17.89	1327.67	50.79	2225.01	657.4	F1	409.3950483	70.36199274	372.9002229	0
A1	A	1	Control	121	T6	FF.03	22DPA	489	3123.06	8.67	450.85	23.53	1194.12	288.49	F9	725.0013475	33.0778677	871.7036417	0
A1	A	1	Control	164	T5	FR.01	42DPA	493	1597.7	5.16	237.69	34.63	578.74	84.92	F17	723.1068785	27.73666946	970.4150597	0
A1	A	1	Control	353	T5	FR.04	55DPA	1019	779.85	17.04	373.57	NA	395.51	153.66	F23	801.7624659	27.50853514	1103.460076	4.104895086
A2	A	2	Control	355	T5	FR.04	55DPA	1023	635.13	9.08	300.12	NA	338.39	100.22	F23	801.7624659	27.50853514	1103.460076	4.104895086
A2	A	2	Control	413	T7	FR.02	47DPA	1139	1073.54	10.4	185.82	20.02	336.79	113.81	F20	965.3261679	31.79354234	1155.474234	3.128774247
A3	A	3	Control	117	T7	FF.01	07DPA	401	3698.58	NA	987.72	33.27	2091.11	637.95	F1	409.3950483	70.36199274	372.9002229	0
A4	A	4	Control	158	T5	FF.04	35DPA	481	1702.19	24.39	276.28	NA	769.92	214.97	F15	751.8424341	23.0485025	948.7250199	0
A5	A	5	Control	109	T7	FF.03	22DPA	385	1767.41	2.97	351.52	43.95	881.43	174.81	F8	651.3230882	45.83076382	651.0915378	0
A5	A	5	Control	154	T7	FF.02	15DPA	433	3415.32	19.08	911.21	64.73	2049.42	296.3	F6	619.3250882	45.83076382	651.0915378	0
A6	A	6	Control	31	T6	FF.01	08DPA	109	5416.24	35.57	1833.13	168.54	2779.48	869.08	F3	356.3376494	48.89175215	353.455502	0
A6	A	6	Control	179	T7	FF.03	28DPA	523	1438.53	11.84	348.65	6.41	494.48	84.64	F11	966.8603372	27.47090189	1062.996002	0
A6	A	6	Control	383	T7	FF.04	34DPA	1079	1419.29	3.81	289.87	20.35	778.23	124.73	F14	873.6366279	22.27264949	1029.943239	0
A6	A	6	Control	425	T7	FR.04	55DPA	1163	348.32	3.47	233.8	98.62	185.62	71.07	F25	1128.526505	29.07120456	1339.54916	5.643624441
A7	A	7	Control	419	T7	FR.03	50DPA	1151	563.66	13.85	245.98	64.19	341.19	87.64	F22	1050.48663	35.66103031	1243.74997	3.503144469
A8	A	8	Control	138	T7	FF.02	15DPA	441	2869.19	22.96	702.99	19.24	1458.06	435.54	F6	619.3250882	45.83076382	651.0915378	0
A8	A	8	Control	143	T6	FF.03	29DPA	451	2273.12	9.52	445.16	NA	982.39	210.03	F12	679.2961808	23.94077746	857.8519232	0
A8	A	8	Control	365	T6	FR.02	47DPA	1043	NA	10.92	304.68	6.12	576.56	145.05	F19	756.6885092	33.90751466	975.4853658	1.544911295
A9	A	9	Control	127	T5	FF.03	27DPA	421	2098.25	13.85	431.06	15.27	841.83	203.7	F10	623.2729217	27.38271169	790.6343093	0
A9	A	9	Control	188	T6	FR.01	42DPA												
A10	A	10	Control	33	T6	FF.01	08DPA												
A14	A	14	Control	99	T5	FF.02	15DPA												
A14	A	14	Control	103	T6	FF.02	15DPA												
A14	A	14	Control	119	T7	FF.01	07DPA												
A14	A	14	Control	119	T7	FF.01	07DPA												
A14	A	14	Control	181	T7	FF.03	28DPA												
A14	A	14	Control	367	T6	FR.02	47DPA												
A15	A	15	Control	97	T5	FF.02	15DPA												
A15	A	15	Control	123	T6	FF.03	22DPA												
A15	A	15	Control	377	T6	FR.04	55DPA												
A15	A	15	Control	423	T7	FR.03	50DPA												
A17	A	17	Control	136	T7	FF.02	15DPA												

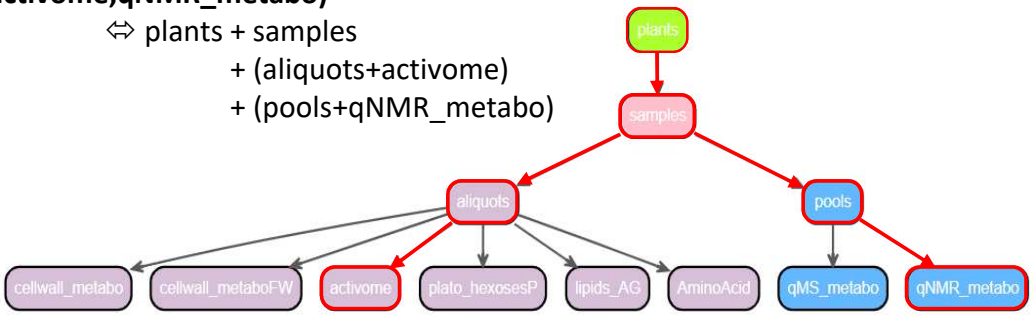
plants samples activome qNMR_metabo

Merging & selection of data subsets

- ⇨ **Avoids lots of data manipulation**
- ⇨ **Facilitates linking both metadata and data for data mining**

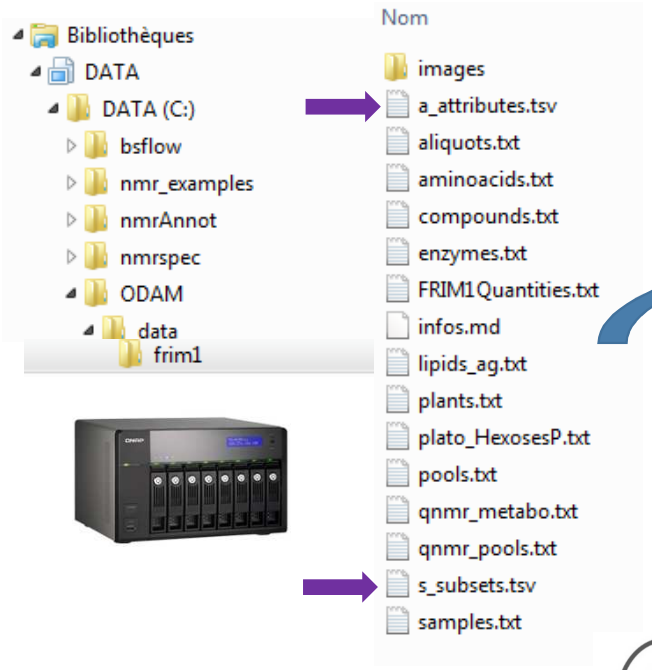
(activome,qNMR_metabo)

⇔ plants + samples
+ (aliquots+activome)
+ (pools+qNMR_metabo)



Example based on FRIM - Fruit Integrative Modelling

Data Emancipation



ODAM

EDTMS

F
A
I NTEROPERABLE
R

With the help of the two metadata files

⇒ Data emancipation regarding Tools

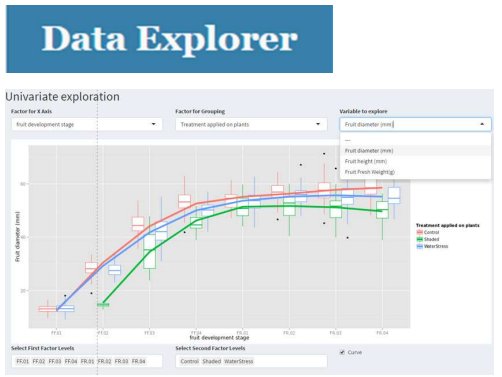
Data ⇔API ⇔ Tools

Data

API

Tools

biostatflow.org



Multiscale deployment
Local / Intranet / Internet

Develop if needed, lightweight tools
- R scripts (Galaxy), lightweight GUI (R shiny)

https://shiny.rstudio.com/
https://plot.ly/dash/

ODAM - Data Explorer

Visual data exploration
a first key step for deeper analyses

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

FRIM1

Example online
https://pmb-bordeaux.fr/dataexplorer/?ds=frim1

ODAM

EDTMS

API

plotly
https://plot.ly/

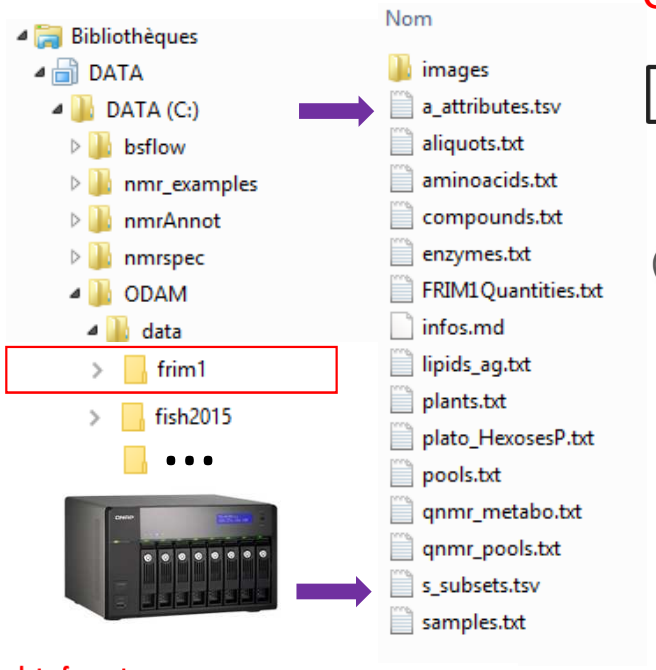
API

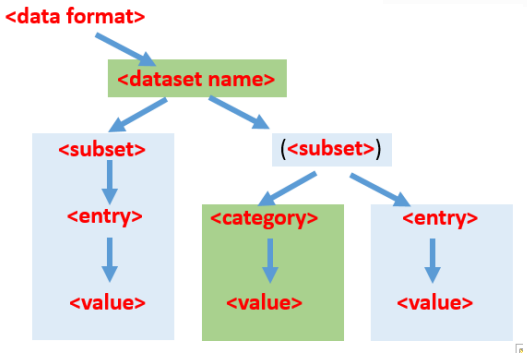
R shiny

ODAM - Data Explorer ≡ frim1

FRIM1

With the help of the two metadata files





ODAM

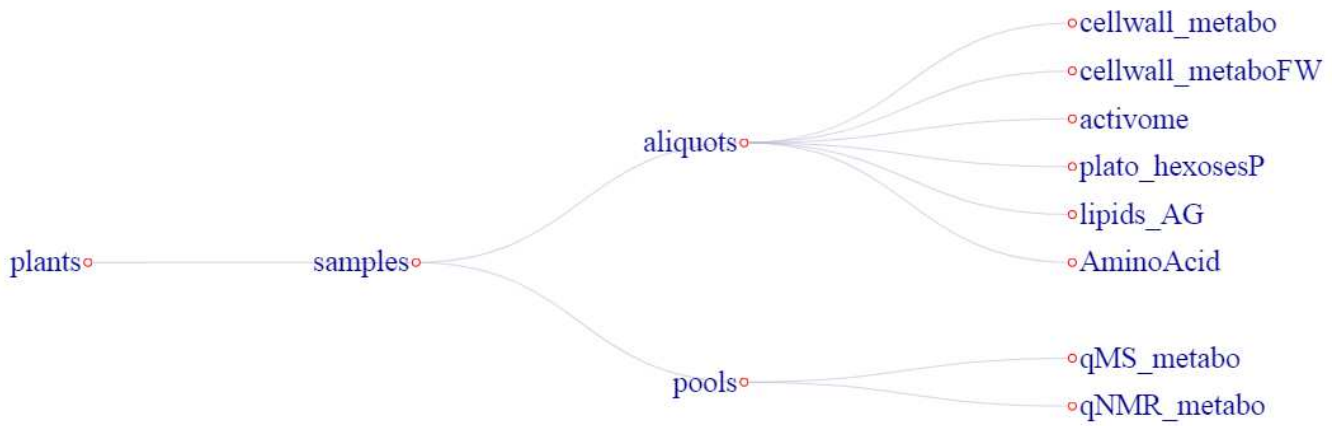
EDTMS

API

Data Subset Information

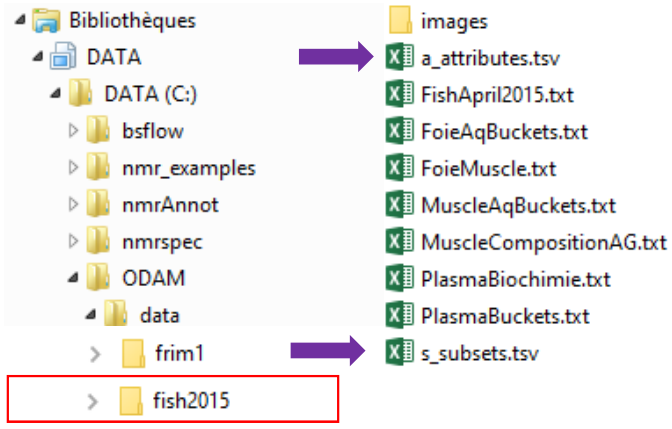
Subset	Description	Identifier	WSEntry	CV_Term
samples	Sample features	SampleID	sample	[OBI_1110046] organ harvesting
cellwall_metabo	Cell wall Compound quantifications	AliquotID	aliquot	[CHEBI_24431] chemical entity
cellwall_metaboFW	Cell Wall Compound quantifications (FW)	AliquotID	aliquot	[CHEBI_24431] chemical entity
activome	Activome Features	AliquotID	aliquot	[CHEBI_24431] chemical entity
qMS_metabo	MS Compounds quantification	PoolID	pool	[CHEBI_24431] chemical entity
qNMR_metabo	NMR Compounds quantification	PoolID	pool	[OBI_1110046] organ harvesting
plato_hexosesP	Hexoses Phosphate	AliquotID	aliquot	[CHEBI_24431] chemical entity
lipids_AG	Lipids AG	AliquotID	aliquot	[CHEBI_24431] chemical entity
AminoAcid	Amino Acids	AliquotID	aliquot	[CHEBI_24431] chemical entity

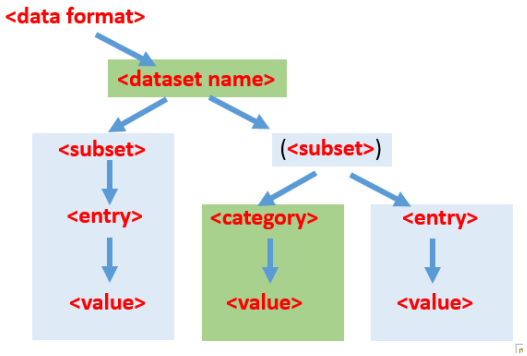
Showing 1 to 9 of 9 entries

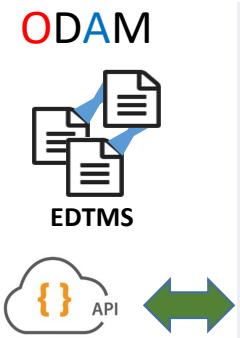


Daniel Jacob – INRA UMR 1332 BFP – Oct 2019

With the help of the two metadata files



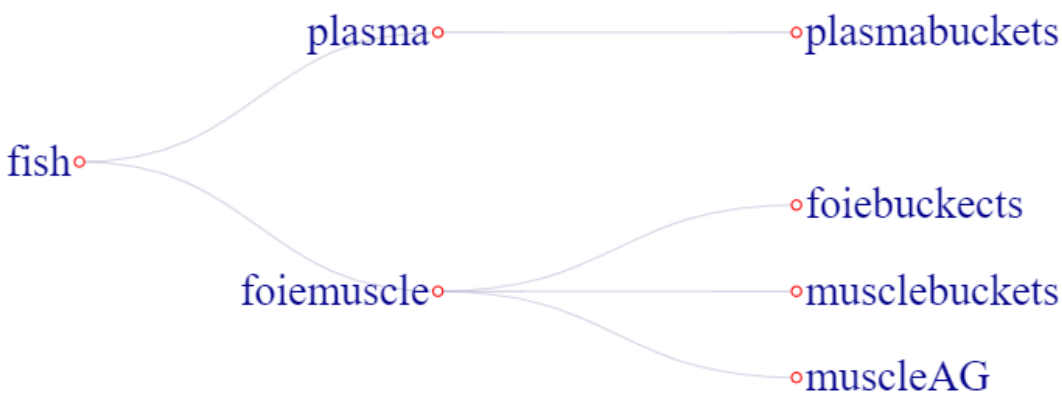


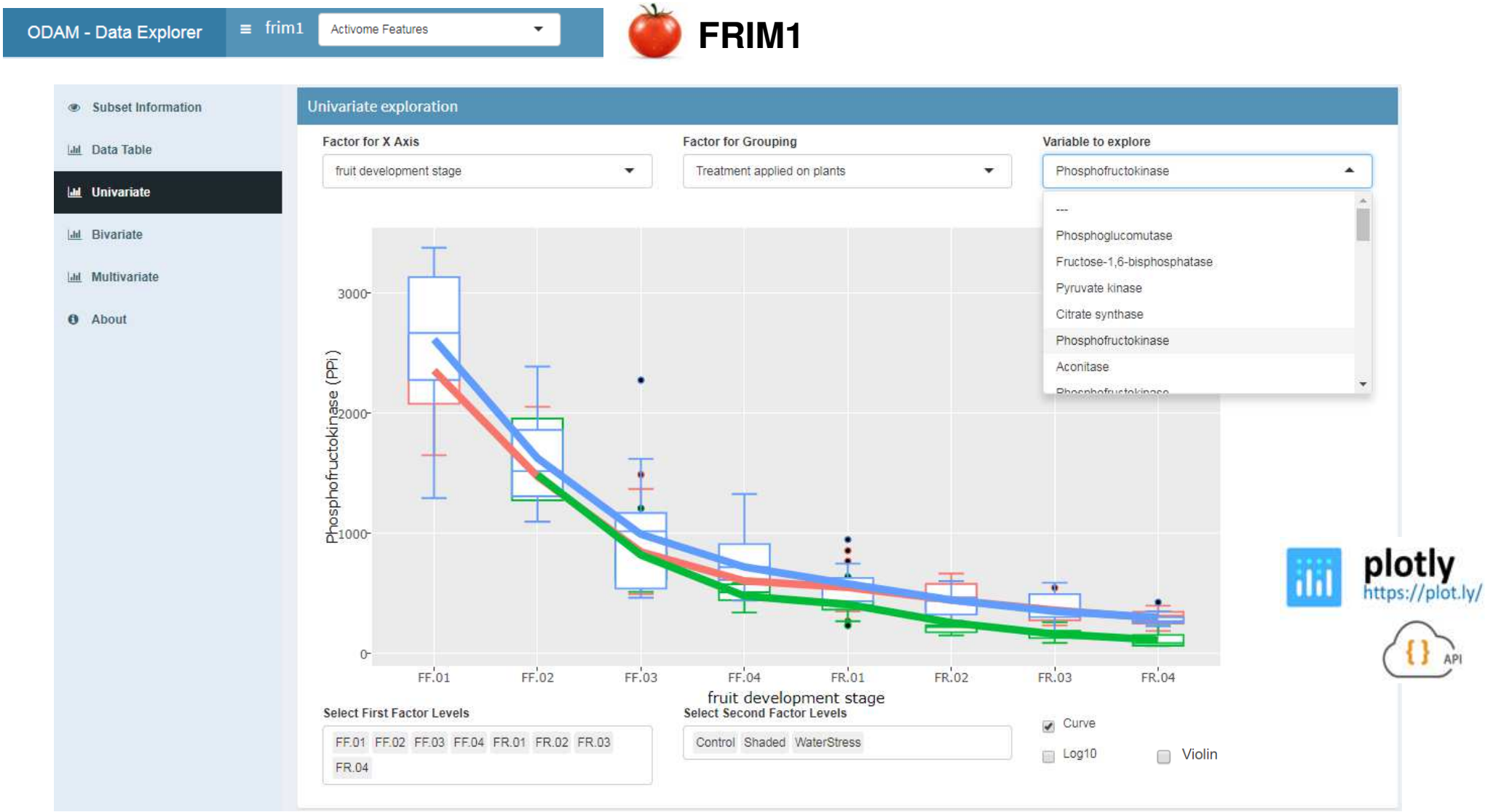


Data Subset Information

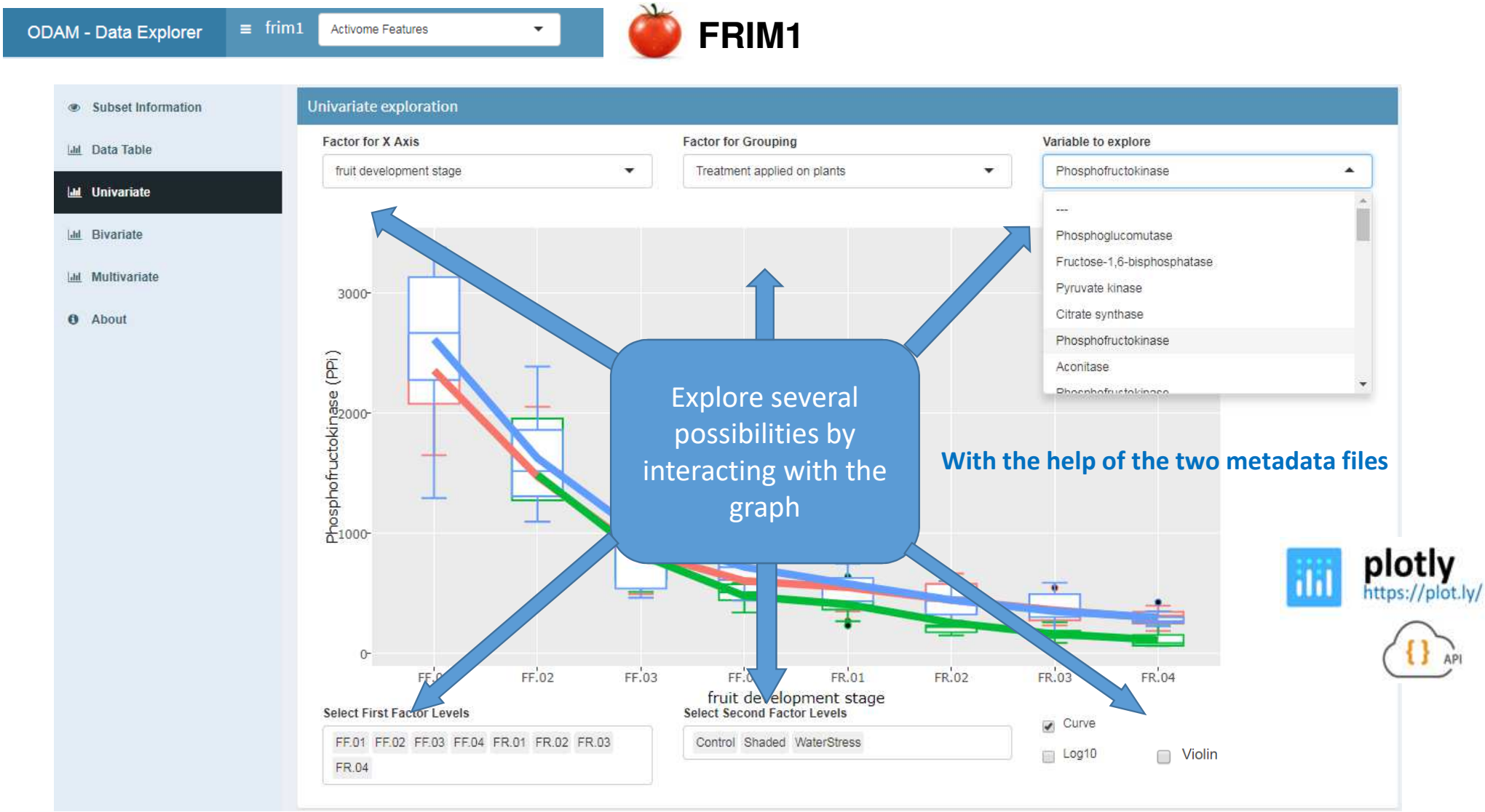
Subset	Description	Identifier	WSEntry	CV_Term
fish	Fish features	Poisson_id	poisson	[null] NULL
plasma	Plasma Biochimie	Plasma_id	plasma	[null] NULL
foiemuscle	Foie Muscle Vol et Poids	Sample_id	sample	[null] NULL
plasmabuckets	Plasma Buckets	Plasma_id		[null] NULL
foiebuckects	Foie Buckets	Faq_id	faq	[null] NULL
musclebuckets	Muscle Buckets	Maq_id	maq	[null] NULL
muscleAG	Compo AG muscles analyses Numea	MAG_id	muscleag	[null] NULL

Showing 1 to 7 of 7 entries



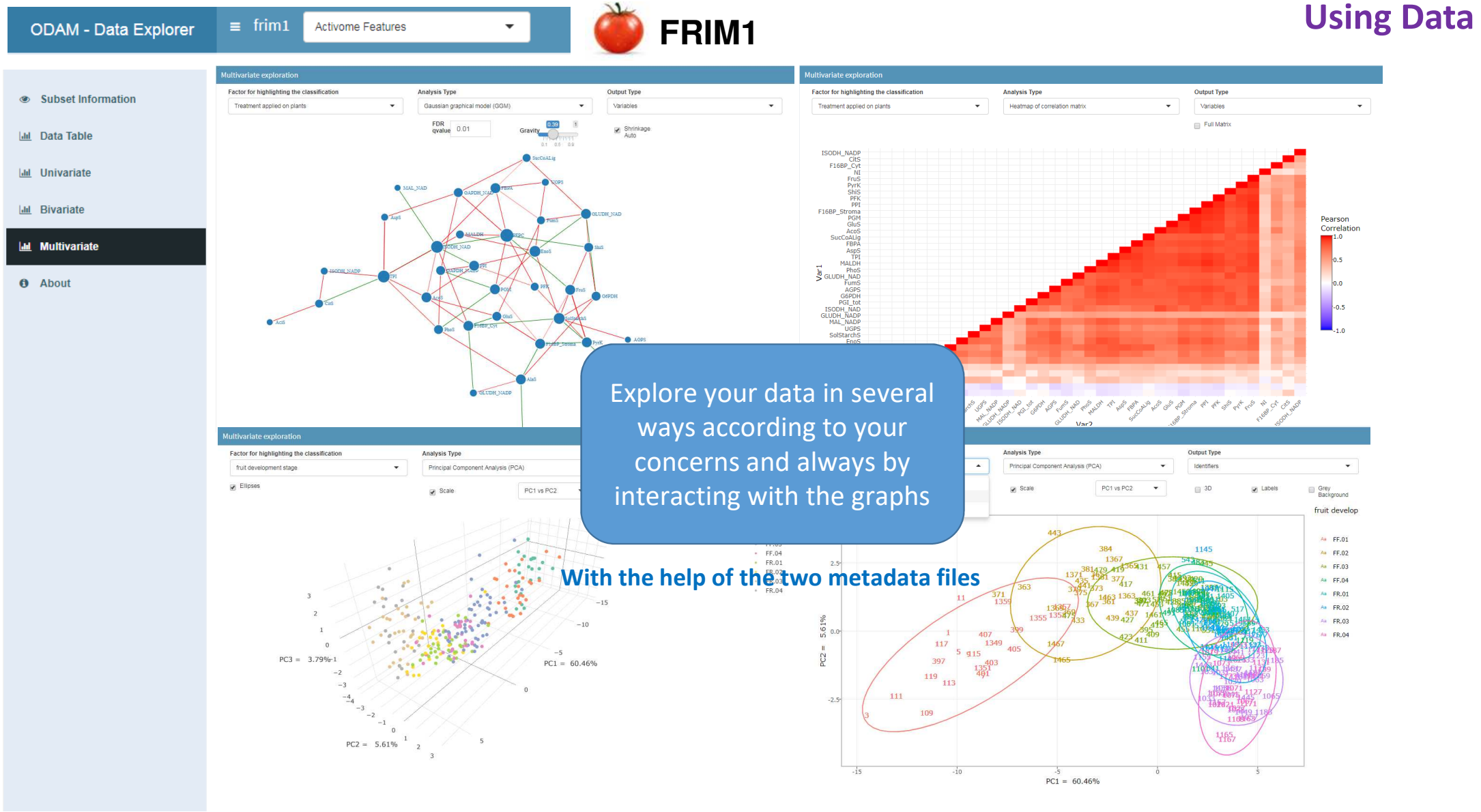

ODAM - Data Explorer
frim1
Activome Features
FRIM1
Subset Information
Data Table
Univariate
Bivariate
Multivariate
About
Univariate exploration
Factor for X Axis
fruit development stage
Factor for Grouping
Treatment applied on plants
Variable to explore
Phosphofructokinase

Phosphoglucomutase
Fructose-1,6-bisphosphatase
Pyruvate kinase
Citrate synthase
Phosphofructokinase
Aconitase
3000
2000
1000
0
Phosphofructokinase (PPi)
FF.01
FF.02
FF.03
FF.04
FR.01
FR.02
FR.03
FR.04
fruit development stage
Select First Factor Levels
FF.01 FF.02 FF.03 FF.04 FR.01 FR.02 FR.03 FR.04
Select Second Factor Levels
Control Shaded WaterStress
Curve
Log10
Violin
plotly
https://plot.ly/
API


ODAM - Data Explorer
frim1
Activome Features
FRIM1
Subset Information
Data Table
Univariate
Bivariate
Multivariate
About
Univariate exploration
Factor for X Axis
fruit development stage
Factor for Grouping
Treatment applied on plants
Variable to explore
Phosphofructokinase

Phosphoglucomutase
Fructose-1,6-bisphosphatase
Pyruvate kinase
Citrate synthase
Phosphofructokinase
Aconitase
Explore several possibilities by interacting with the graph
With the help of the two metadata files
plotly
https://plot.ly/
API
Phosphofructokinase (PPi)
3000
2000
1000
0
FF.02
FF.03
FR.01
FR.02
FR.03
FR.04
fruit development stage
Select First Factor Levels
FF.01 FF.02 FF.03 FF.04 FR.01 FR.02 FR.03 FR.04
Select Second Factor Levels
Control Shaded WaterStress
Curve
Log10
Violin

Using Data



Daniel Jacob – INRA UMR 1332 BFP – Oct 2019

Using Data

ODAM - Data Explorer ≡ frim1 Activome Features

FRIM1

Subset Information
Data Table
Univariate
Bivariate
Multivariate
About

Columns to show:
PlantID
Rank
PlantNum
Treatment
SampleID
Truss
DevStage
FruitAge
HarvestDate
HarvestHour
FruitPosition
FruitDiameter
FruitHeight
FruitFW
DW
AliquotID
PGM
F16BP_Cyt
PyrK
CitS
PPI
AcoS
PFK
F16BP_Stroma
GluS
ISODH_NAD
EnoS
ISODH_NADP
PEPC
FBPA
SucCoALig
MALDH
AlaS
FumS

Copy Excel Search:

Treatmen	DevStage	FruitAge	AliquotID	PGM	F16BP_Cyt	PyrK	CitS	PPI	AcoS	PFK	F16BP_Stroma	GluS	ISODH_NAD
											All		A
Control	FF.01	07DPA	397	4924.26	17.89	1327.67	50.79	2225.01	657.4	160.33	398.38	458.04	495.24
Control	FF.03	22DPA	409	3123.06	8.67	450.85	23.53	1194.12	288.49	125.19	209.7	218.39	188.53
Contr			493	1597.7	5.16	237.69	34.63	578.74	84.92	66.91	72.58	47.22	133.34
Contr												6.47	314.05
Control	FR.04	55DPA										9.9	227.51
Control	FR.02	47DPA										121.14	275.09
Control	FF.03	21DPA											89.87
Control	FF.01	07DPA										368.98	336.88
Control	FR.02	47DPA										21.28	296.93
Control	FR.03	50DPA										9.47	306.54
Control	FF.04	35DPA										70.67	197.18
Control	FF.03	22DPA										82.19	135.55
Control	FF.02	15DPA										265.13	180.87
Control	FF.01	08DPA										614.79	660.68
Control	FF.03	28DPA										34.81	96.16
Control	FF.04	34DPA										55.6	122.12
Control	FR.04	55DPA										15.32	161.75
Control	FF.03	30DPA										9.84	122.83
Control	FR.03	50DPA										28.09	266.62
Control	FF.02	15DPA										394.7	176.69
Control	FF.03	29DPA										69.56	121.26
Control	FR.02	47DPA										26.46	222.35
Control	FF.03	27DPA										90.63	137.55
Control	FR.01	42DPA										25.77	114.93
Control	FF.01	08DPA										526.31	437.56

Showing 1 to 25 of 255 entries

Export as …

ODAM - Data Explorer.xlsx - Excel

FICHIER ACCUEIL Perso INSERTION MISE EN PAGE FORMULES DONNÉES RÉVISION AFFICHAGE DÉVELOPPEUR

Calibri 11 Renvoyer à la ligne automatiquement Standard

Fusionner et centrer

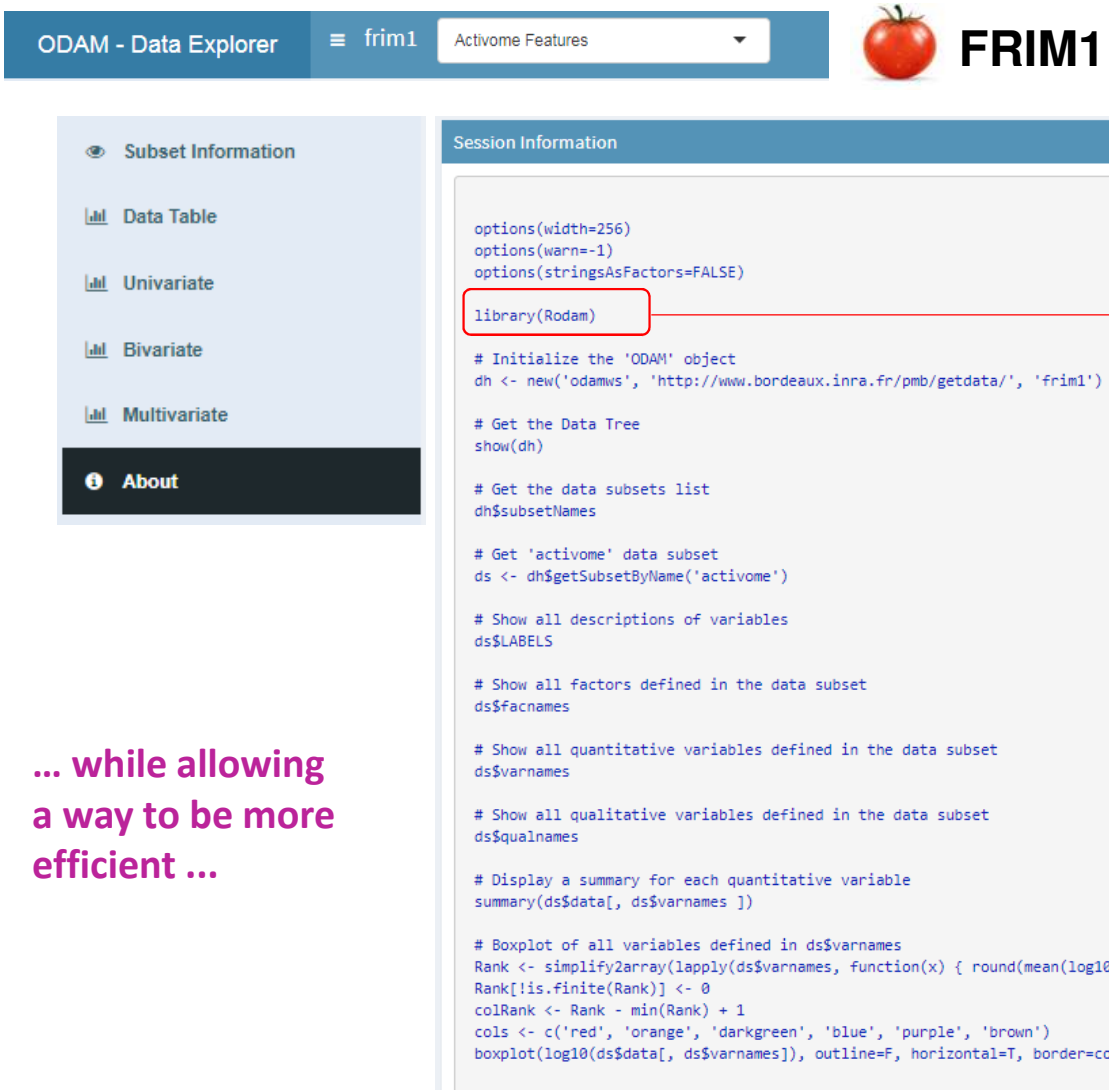
Police Alignement Nombre

A1 Treatment

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Treatment	DevStag	FruitAge	AliquotID	PGM	F16BP_Cy	PyrK	CitS	PPI	AcoS	PFK	F16BP_Strom	GluS	ISODH_NAD	
2	Control	FF.01	07DPA	397	4924.3	17.89	1327.7	50.79	2225	657.4	160.3	398.38	458.04	495.24	
3	Control	FF.03	22DPA	409	3123.1	8.67	450.85	23.53	1194.1	288.49	125.2	209.7	218.39	188.53	
4	Control	FR.01	42DPA	493	1597.7	5.16	237.69	34.63	578.74	84.92	66.91	72.58	47.22	133.34	
5	Control	FR.04	55DPA	1019	779.85	17.04	373.57		395.51	153.66	66.41	37.27	6.47	314.05	
6	Control	FR.04	55DPA	1023	635.13	9.08	300.12		338.39	100.22	64.27	33.65	9.9	227.51	
7	Control	FR.02	47DPA	1139	1073.5	10.4	185.82	20.02	336.79	113.81	53.05	43.88	121.14	275.09	
8	Control	FF.03	21DPA	1373	1446.7	6.44	259.53		493.07	102.94	68.73	85.44		89.87	
9	Control	FF.01	07DPA	401	3698.6		987.72	33.27	2091.1	637.95	118.8	247.34	368.98	336.88	
10	Control	FR.02	47DPA	1421	1337	6.91	226.24		508.56	110.26	86	82.63	21.28	296.93	
11	Control	FR.03	50DPA	1437	484.66	13.28	172.04	24.19	348.75	79.06	49.86	41.97	9.47	306.54	
12	Control	FF.04	35DPA	481	1702.2	24.39	276.28		769.92	214.97	88.58	130.93	70.67	197.18	
13	Control	FF.03	22DPA	385	1767.1	11.97	351.52	43.95	881.43	174.81	85.89	127.57	82.19	135.55	
14	Control	FF.02	15DPA	433	3115.5	19.08	911.21	64.73	2049.4	296.3	117	241.41	265.13	180.87	
15	Control	FF.01	08DPA	109	5416.2	35.57	1833.1	168.5	2779.5	869.08	197.7	494.3	614.79	660.68	
16	Control	FF.03	28DPA	523	1438.5	11.84	348.65	6.41	494.48	84.64	70.67	33.81	34.81	96.16	
17	Control	FF.04	34DPA	1079	1419.3	3.81	289.87	20.35	778.23	124.73	68.16	57.51	55.6	122.12	
18	Control	FR.04	55DPA	1163	348.32	3.47	233.8	98.62	185.62	71.07	33.76	5.28	15.32	161.75	
19	Control	FF.03	30DPA	1389	1069.8	9.14	189.82	34.23	538.35	100.75	54.16	58.85	9.84	122.83	
20	Control	FR.03	50DPA	1151	563.66	13.85	245.98	64.19	341.19	87.64	44.91	49.19	28.09	266.62	
21	Control	FF.02	15DPA	441	2869.2	22.96	702.99	19.24	1458.1	435.54	129.4	200.16	394.7	176.69	

As far as possible, **keep the old way of using the scientist's worksheets …**

Using Data



The Comprehensive R Archive Network
https://cran.r-project.org

The R package ***Rodam***

… while allowing a way to be more efficient …

Retrieving Data within R

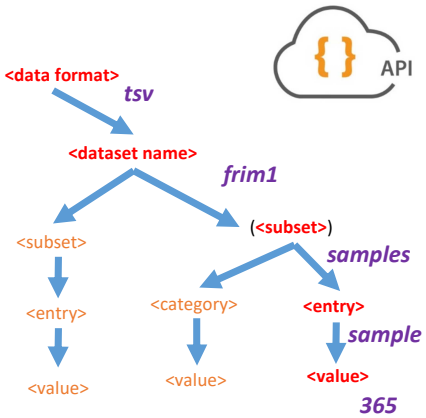
CRAN 0.1.4

The R package
Rodam

The Comprehensive R Archive Network
https://cran.r-project.org/web/packages/Rodam/index.html

```
library(Rodam)

#Get the metadata of the 'friml' dataset
dh <- new('odamws', 'http://pmb-bordeaux.fr/getdata/', 'friml')
```

```
                    levelName SetID Identifier WSEntry                                Description Count
plants                              1    PlantID   plant                        Plant features   552
 °--samples                         2   SampleID  sample                        Sample features  1287
     ¦--aliquots                    3 AliquotID aliquot                      Aliquots features  2571
     ¦   ¦--cellwall_metabo         4 AliquotID aliquot      Cell wall Compound quantifications   343
     ¦   ¦--cellwall_metaboFW       5 AliquotID aliquot Cell Wall Compound quantifications (FW)   343
     ¦   ¦--activome                6 AliquotID aliquot                     Activome Features  1260
     ¦   ¦--plato_hexosesP         10 AliquotID aliquot                     Hexoses Phosphate  1260
     ¦   ¦--lipids_AG              11 AliquotID aliquot                             Lipids AG   272
     ¦   °--AminoAcid              12 AliquotID aliquot                             Amino Acids 342
     °--pools                       7     PoolID    pool                 Pools of remaining pools 948
         ¦--qMS_metabo              8     PoolID    pool          MS Components quantification   399
         °--qNMR_metabo             9     PoolID    pool         NMR Components quantification   948
```




```
# get data from 'samples' subset with a constraint
dh$getDataByName('samples','sample/365')
```

	PlantID	Rank	PlantNum	Treatment	SampleID	Truss	DevStage	FruitAge	HarvestDate	HarvestHour	FruitPosition	FruitDiameter	FruitHeight	FruitFW
1	E35	E	311	Control	365	T6	FR.02	47DPA	40423	0.5	5	55.46	48.98	83.32
2	A17	A	17	Control	365	T6	FR.02	47DPA	40423	0.5	3	56.59	47.77	82.02
3	A8	A	8	Control	365	T6	FR.02	47DPA	40423	0.5	5	55.11	44.90	71.82
4	D3	D	210	Control	365	T6	FR.02	47DPA	40423	0.5	5	49.28	44.35	58.28
5	H11	H	356	Control	365	T6	FR.02	47DPA	40423	0.5	6	46.68	38.69	49.25

GET http://pmb-bordeaux.fr/getdata/tsv/frim1/(samples)/sample/365

Example based on FRIM - Fruit Integrative Modelling

Retrieving Data within R

CRAN 0.1.4

data and metadata are all available and accessible by scripting languages(R, API)

Merge the activome data **AND** the metabolome data acquire on the same samples

```
### Load the R ODAM package
library(Rodam)

### Get the merged data of two data subsets based on their common identifiers
dh <- new('odamws', wsURL='https://pmb-bordeaux.fr/getdata/', dsname='friml')
setNameList <- c("activome", "qNMR_metabo" )
dsMerged <- dh$getSubsetByName(setNameList)
```

API

```
> dim(dsMerged$data)
[1] 948  83

> dsMerged$LABELS
        Subset     Attribute                        Description       Type
1      samples      SampleID          Pool of several harvests Identifier
2       plants     Treatment        Treatment applied on plants     Factor
3      samples      DevStage           fruit development stage     Factor
4      samples      FruitAge                   fruit age (dpa)     Factor
5     activome           PGM                  Phosphoglucomutase   Variable
6     activome      F16BP_Cyt  Fructose-1,6-bisphosphatase (cyt)   Variable
7     activome           PyrK                  Pyruvate kinase   Variable
8     activome           CitS                  Citrate synthase   Variable
9     activome           PPI          Phosphofructokinase (PPi)   Variable
10    activome           AcoS                       Aconitase   Variable
 ...

43 qNMR_metabo       glucose            glucose (nMol / mg MS)   Variable
44 qNMR_metabo    saccharose         saccharose (nMol / mg MS)   Variable
45 qNMR_metabo      fructose           fructose (nMol / mg MS)   Variable
46 qNMR_metabo     galactose          galactose (nMol / mg MS)   Variable
47 qNMR_metabo       mannose            mannose (nMol / mg MS)   Variable
48 qNMR_metabo      rhamnose           rhamnose (nMol / mg MS)   Variable
49 qNMR_metabo       acetate            acetate (nMol / mg MS)   Variable
50 qNMR_metabo  chlorogenate       chlorogenate (nMol / mg MS)   Variable
 ...
```

data

metadata

The Comprehensive R Archive Network
https://cran.r-project.org

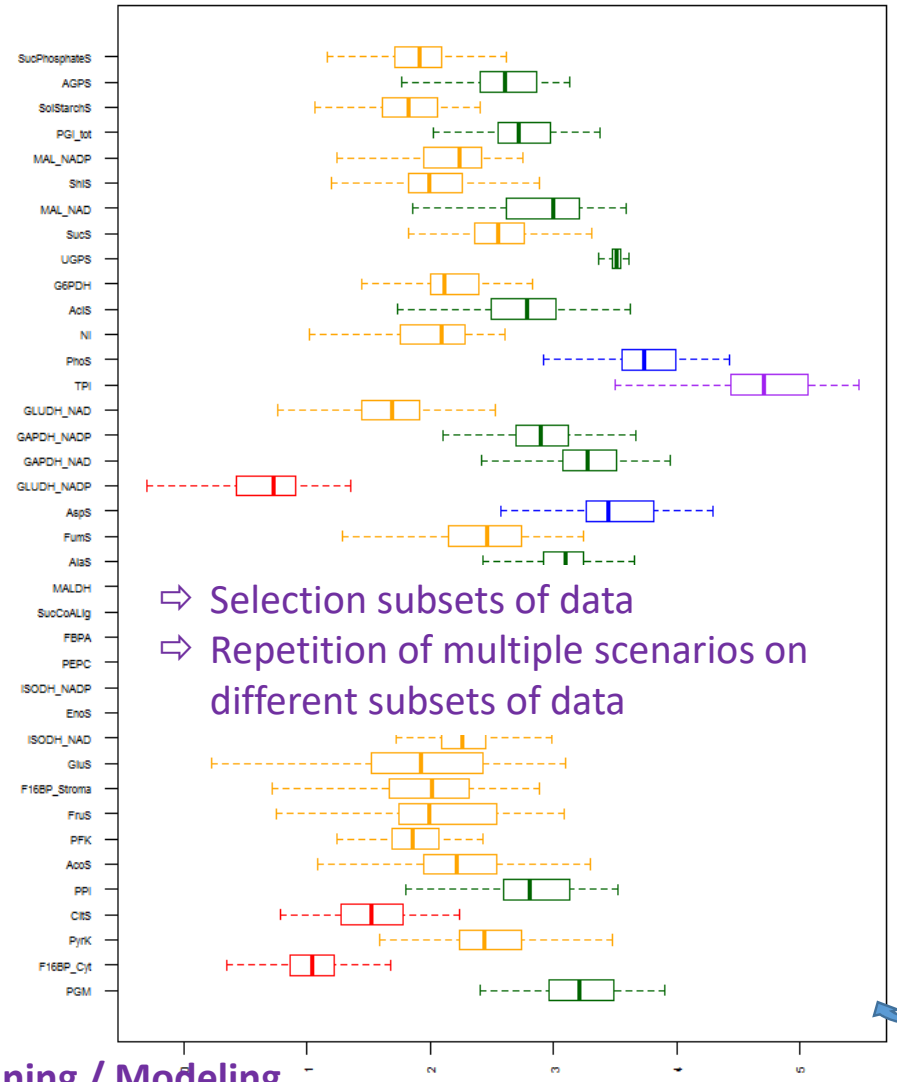
The R package ***Rodam***

Example based on FRIM - Fruit Integrative Modelling

Retrieving Data within R

CRAN 0.1.4

data and metadata are all available and accessible by scripting languages(R, API)



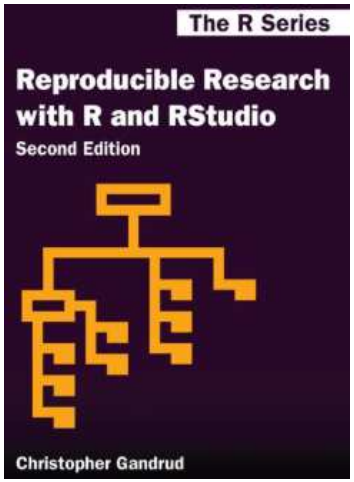
⇨ Selection subsets of data

⇨ Repetition of multiple scenarios on different subsets of data

```
library(Rodam)

#Get the metadata of the 'friml' dataset
dh <- new('odamws', 'http://pmb-bordeaux.fr/getdata/', 'friml')

# get 'activome' data subset
ds <- dh$getSubsetByName('activome')

ds$samples   # Show the identifier defined in the data subset
[1] "AliquotID"

ds$facnames   # Show all factors defined in the data subset
[1] "Treatment" "DevStage"  "FruitAge"

ds$varnames   # Show all quantitative variables defined in the data subset
 [1] "PGM"          "F16BP_Cyt"    "PyrK"         "CitS"
 [5] "PPI"          "AcoS"         "PFK"          "FruS"
 [9] "F16BP_Stroma" "GluS"         "ISODH_NAD"    "EnoS"
[13] "ISODH_NADP"   "PEPC"         "FBPA"         "SucCoALig"
[17] "MALDH"        "AlaS"         "FumS"         "AspS"
[21] "GLUDH_NADP"   "GAPDH_NAD"    "GAPDH_NADP"   "GLUDH_NAD"
[25] "TPI"          "PhoS"         "NI"           "AciS"
[29] "G6PDH"        "UGPS"         "SucS"         "MAL_NAD"
[33] "ShiS"         "MAL_NADP"     "PGI_tot"      "SolStarchS"
[37] "AGPS"         "SucPhosphateS"

ds$qualnames # Show all qualitative variables defined in the data subset
[1] "Rank"  "Truss"

# Boxplot of all variables defined in ds$varnames
par(mar=c(2,8,1,2))
boxplot(log10(ds$data[, ds$varnames]), horizontal=T, outline=F, las=2, cex.axis=0.7)
```

Data mining / Modeling

Reproducible Research … with R and RStudio



Christopher Gandrud (2015)

Chap II **Data Gathering and Storage** (70 pages out of 300)

II. 6 - **Gathering Data with R**

"How you gather your data directly impacts how reproducible your research will be.

If all of your data gathering steps are tied together by your source code, then independent researchers (and you) can more easily regather the data"

II. 7 - **Preparing Data for Analysis**

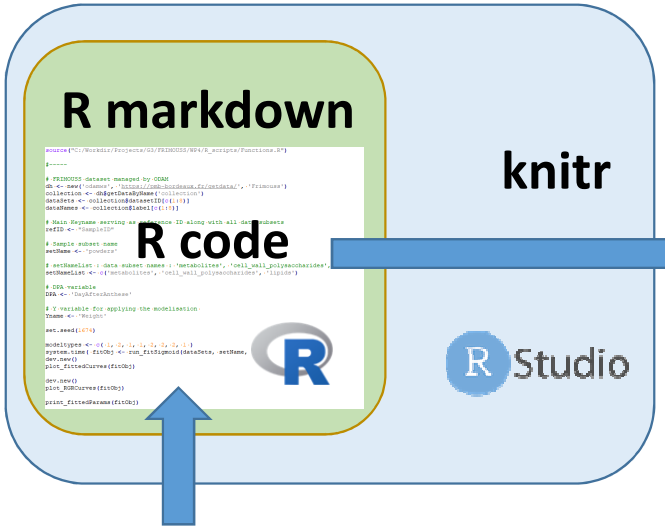
"Once we have gathered the raw data that we want to include in our statistical analyses we generally need to clean it up so that it can be merged into a single data file."

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

This is exactly what the ODAM framework aims to answer in a normalized way the easier and faster as possible

https://github.com/christophergandrud/Rep-Res-Book
https://englianhu.files.wordpress.com/2016/01/reproducible-research-with-r-and-studio-2nd-edition.pdf

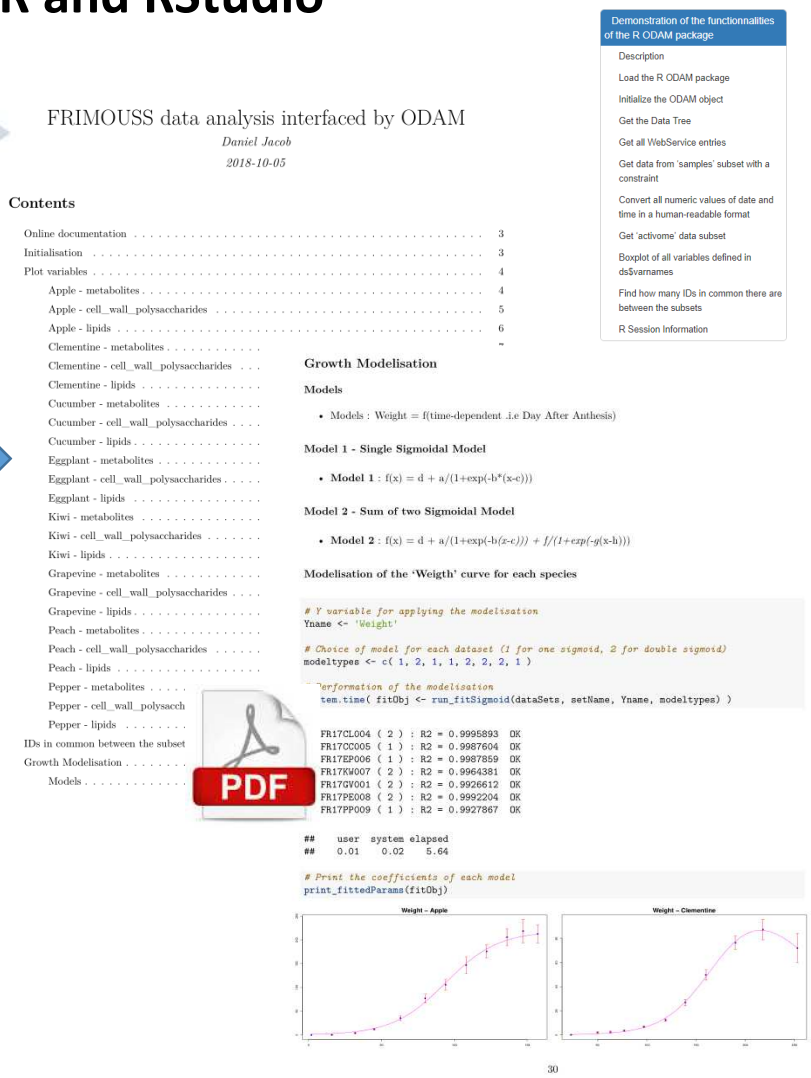
Reproducible Research … with R and RStudio



The R package
Rodam

CRAN 0.1.4

ODAM Framework



FRIMOUSS data analysis interfaced by ODAM

Daniel Jacob

2018-10-05

Contents

Growth Modelisation

Models

- Models : Weight = f(time-dependent .i.e Day After Anthesis)

Model 1 - Single Sigmoidal Model

- **Model 1** : $f(x) = d + a/(1+\exp(-b*(x-c)))$

Model 2 - Sum of two Sigmoidal Model

- **Model 2** : $f(x) = d + a/(1+\exp(-b(x-c))) + f/(1+\exp(-g(x-h)))$

Modelisation of the 'Weigth' curve for each species

```
# Y variable for applying the modelisation
Yname <- 'Weight'

# Choice of model for each dataset (1 for one sigmoid, 2 for double sigmoid)
modeltypes <- c( 1, 2, 1, 1, 2, 2, 2, 1 )

# Performation of the modelisation
system.time( fitObj <- run_fitSigmoid(dataSets, setName, Yname, modeltypes) )
```

```
FR17CL004 ( 2 ) : R2 = 0.9995893  OK
FR17CC005 ( 1 ) : R2 = 0.9987604  OK
FR17EP006 ( 1 ) : R2 = 0.9987859  OK
FR17KW007 ( 2 ) : R2 = 0.9964381  OK
FR17GV001 ( 2 ) : R2 = 0.9926612  OK
FR17PE008 ( 2 ) : R2 = 0.9992204  OK
FR17PP009 ( 1 ) : R2 = 0.9927867  OK

##    user  system elapsed
##    0.01    0.02    5.64
```

```
# Print the coefficients of each model
print_fittedParams(fitObj)
```


30

Demonstration of the functionnalities of the R ODAM package

Description

Load the R ODAM package

Initialize the ODAM object

Get the Data Tree

Get all WebService entries

Get data from 'samples' subset with a constraint

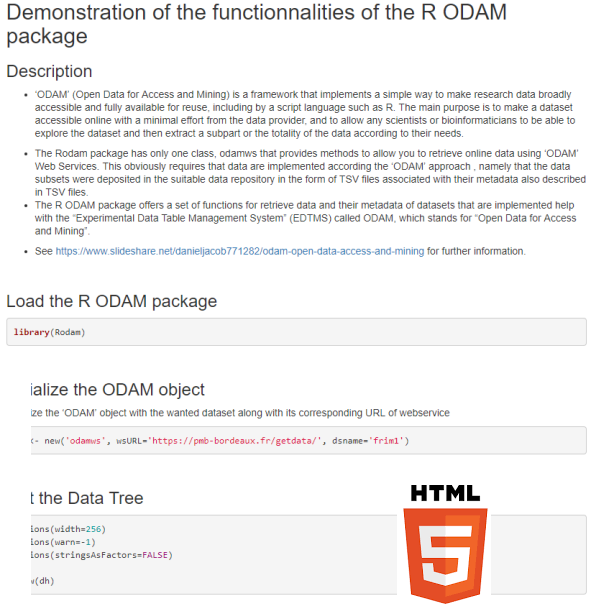
Convert all numeric values of date and time in a human-readable format

Get 'activome' data subset

Boxplot of all variables defined in ds$varnames

Find how many IDs in common there are between the subsets

R Session Information



Demonstration of the functionnalities of the R ODAM package

Description

- 'ODAM' (Open Data for Access and Mining) is a framework that implements a simple way to make research data broadly accessible and fully available for reuse, including by a script language such as R. The main purpose is to make a dataset accessible online with a minimal effort from the data provider, and to allow any scientists or bioinformaticians to be able to explore the dataset and then extract a subpart or the totality of the data according to their needs.
- The Rodam package has only one class, odamws that provides methods to allow you to retrieve online data using 'ODAM' Web Services. This obviously requires that data are implemented according the 'ODAM' approach , namely that the data subsets were deposited in the suitable data repository in the form of TSV files associated with their metadata also described in TSV files.
- The R ODAM package offers a set of functions for retrieve data and their metadata of datasets that are implemented help with the "Experimental Data Table Management System" (EDTMS) called ODAM, which stands for "Open Data for Access and Mining".
- See https://www.slideshare.net/danieljacob771282/odam-open-data-access-and-mining for further information.

Load the R ODAM package

```
library(Rodam)
```

ialize the ODAM object

ze the 'ODAM' object with the wanted dataset along with its corresponding URL of webservice

```
(- new('odamws', wsURL='https://pmb-bordeaux.fr/getdata/', dsname='frim1')
```

t the Data Tree

```
ions(width=256)
ions(warn=-1)
ions(stringsAsFactors=FALSE)

(dh)
```

https://rmarkdown.rstudio.com/authoring_quick_tour.html

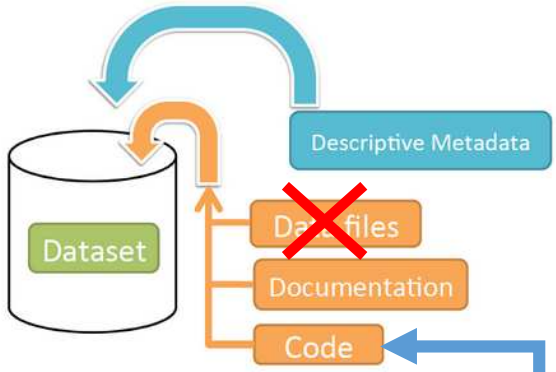
Data Dissemination

FINDABLE
A
I
R

doi:10.15454/95JUTK

The Dataverse Project

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



Container for your data, documentation, and code.

https://data.inra.fr/

FRIM - Fruit Integrative Modelling Version 1.1

Bénard, Camille; Bernillon, Stéphane; Biais, Benoit; Maucourt, Mickael ; Ballias, Patricia; Deborde, Catherine; Colombié, Sophie; Cabasson, Cécile; Jacob, Daniel; Gibon, Yves; Moing, Annick, 2018, "FRIM - Fruit Integrative Modelling", https://doi.org/10.15454/95JUTK, Portail Data Inra, V1

Cite Dataset

Learn about Data Citation Standards.

Description	The project aimed to build a virtual tomato fruit that enables the prediction of metabolite levels given genetic and environmental inputs, by an iterative process between laboratories which combine expertise in fruit biology, ecophysiology, theoretical and experimental biochemistry, and biotechnology.
Subject	Computer science; Information management; Omics; Plant Health and Pathology
Related Publication	Bénard C., Bernillon S., Biais B., Osorio S., Maucourt M., Ballias P., Deborde C., Colombié S., Cabasson C., Jacob D., Vercambre G., Gautier H., Rolin D., Génard M., Fernie A., Gibon Y., Moing A. 2015 Metabolomic profiling in tomato reveals diel compositional changes in fruit affected by source-sink relationships. Journal of Experimental Botany Vol. 66, No. 11 pp. 3391–3404 doi: 10.1093/jxb/erv151
Link to data	https://pmb-bordeaux.fr/dataexplorer/?ds=frim1

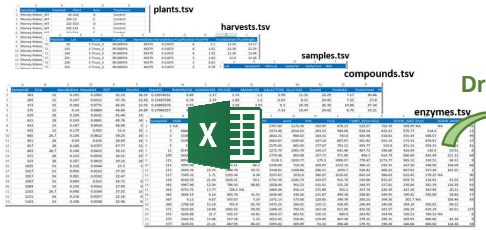
R scripts (Rmd)
Jupyter Notebook (ipynb) *If applicable*

ODAM Framework Overview

Promote good practices → **Provide services**

ODAM
Open Data for Access and Mining

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

data **push**

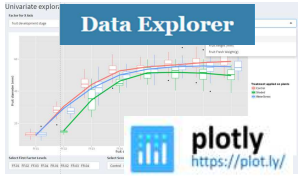
https://hub.docker.com/r/odam/getdata/

GET

Data can be downloaded, explored and mined

F **A** **I**NTEROPERABLE **R**

Multi-criteria exploration & interactive visualization



Link

The Dataverse Project

FINDABLE **A**CCESSIBLE **I** **R**E-USABLE

https://data.inra.fr

Allow users to gain efficiency <u>where they would like to gain efficiency</u>

Merging & selection of data subsets

Export

Repetition of complex treatments according to very varied parameters

API

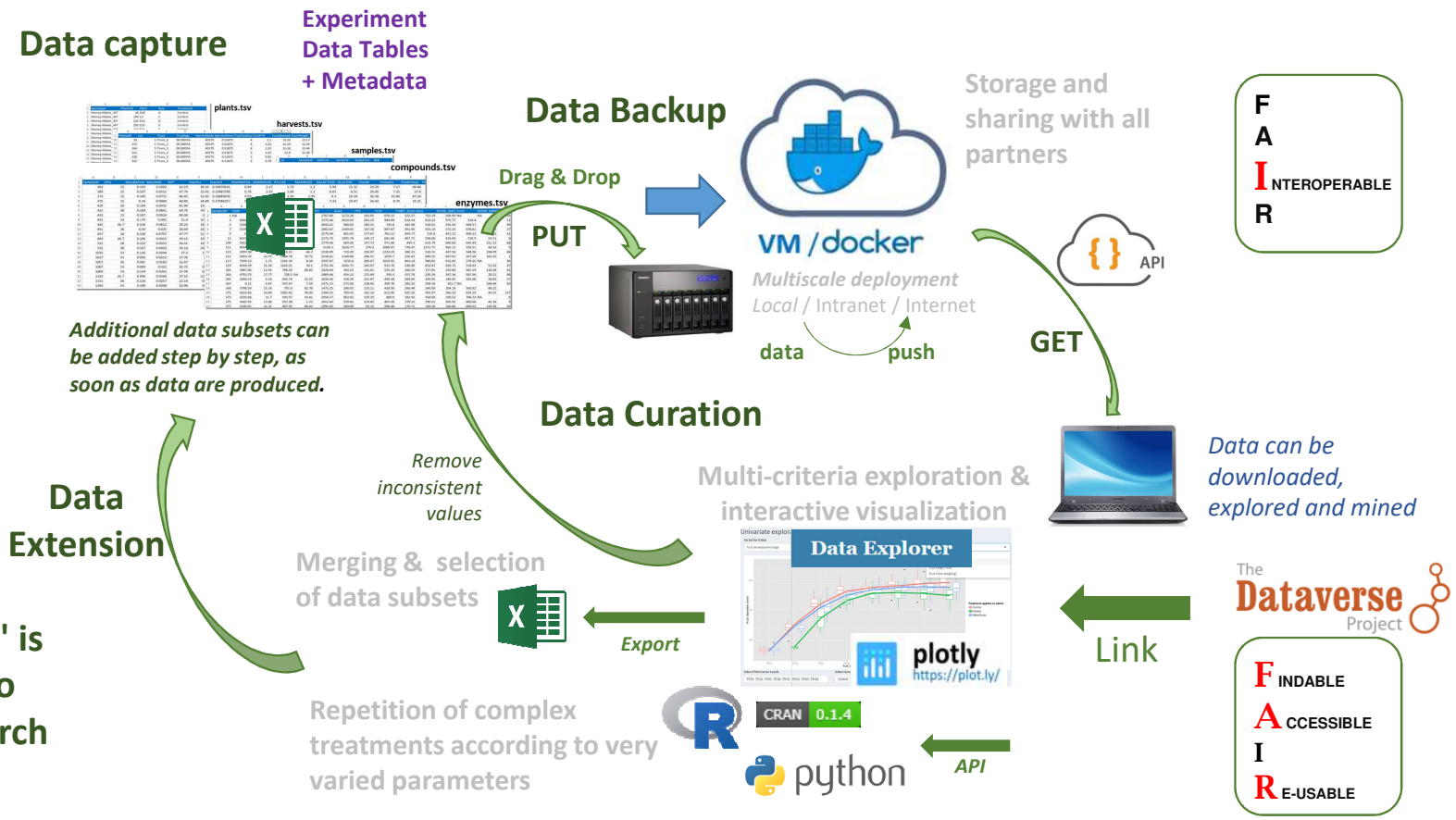
Develop if needed, lightweight tools
- R/Python scripts (Galaxy), lightweight GUI (R shiny)
- Python (Plotly Dash)

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

ODAM Framework Overview

ODAM
Open Data for Access and Mining

Data Lifecycle



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

Data is “keep alive” in the cycle

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019

Advantages of this approach

data sharing & data availability

- **The array of the "plants"** may be created even before planting the seeds.
- Similarly, **the array of the "harvests"** can be created as soon as the harvests are done, and this before any analysis.
- Thus, these arrays are **generated only once in the project** and we can set up the sharing soon the seed planting. Then **each analysis comes to complement the set of data as soon as they produce** their own sub-dataset.
- **data are accessible to everyone** as soon as they are produced,

identifiers centrally managed

- data are archived and compiled, so that it becomes **useless to proceed a laborious investigation to find out who possesses the right identifiers**, etc.

facilitate the subsequent publication of data

- data are already readily available online by web API,
- But nothing prevents to take this data to fill in existing a data repository, by adjoining more elaborate annotations such as Dataverse.

Advantages of this approach

minimal effort, maximum efficiency

Format the data

- Based on TSV: choice to **keep the good old way of scientist to use worksheets**, thus *i*) using the same tool for both data files and metadata definition files, *ii*) no programmatic skill are required

Give an access through a web services layer

- based on current standards (REST)

Use existing tools

- Spreadsheets, R studio, Spyder, Jupyterlab, BioStatFlow, …

Develop if needed, lightweight tools

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

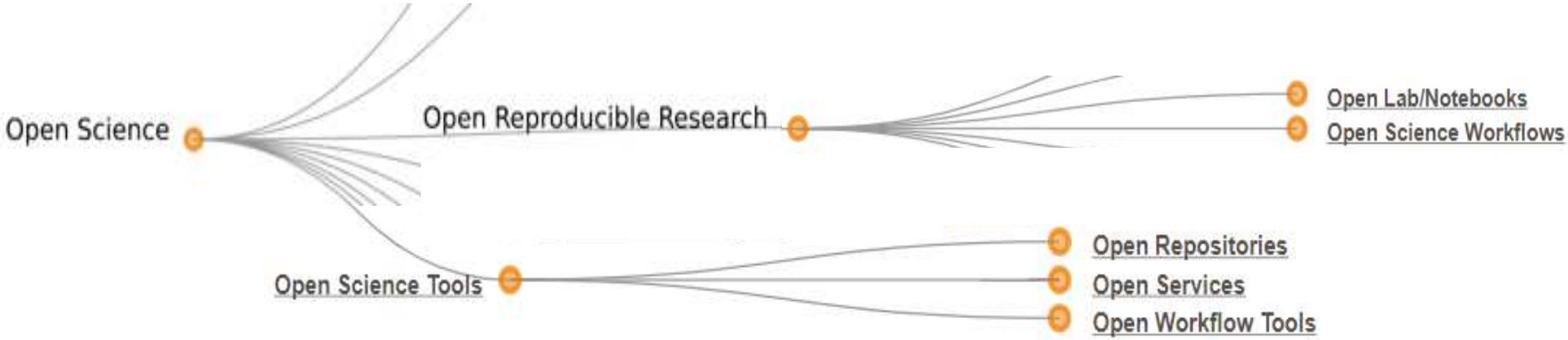






Tools for data science

Reproducible Research



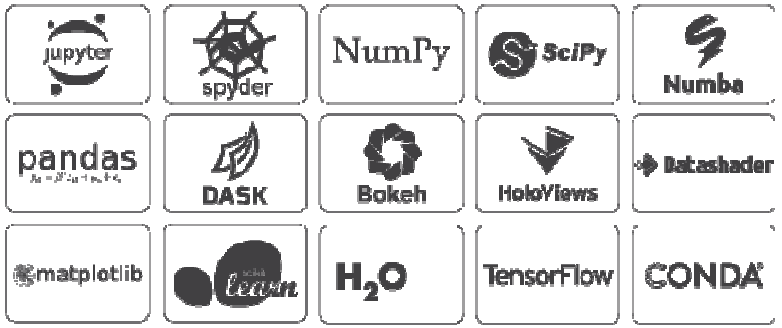
Tools for data science

Anaconda Distribution

Windows | macOS | Linux

https://www.anaconda.com/distribution/

The World's Most Popular Python/R Data Science Platform



+ IRkernel

RStudio

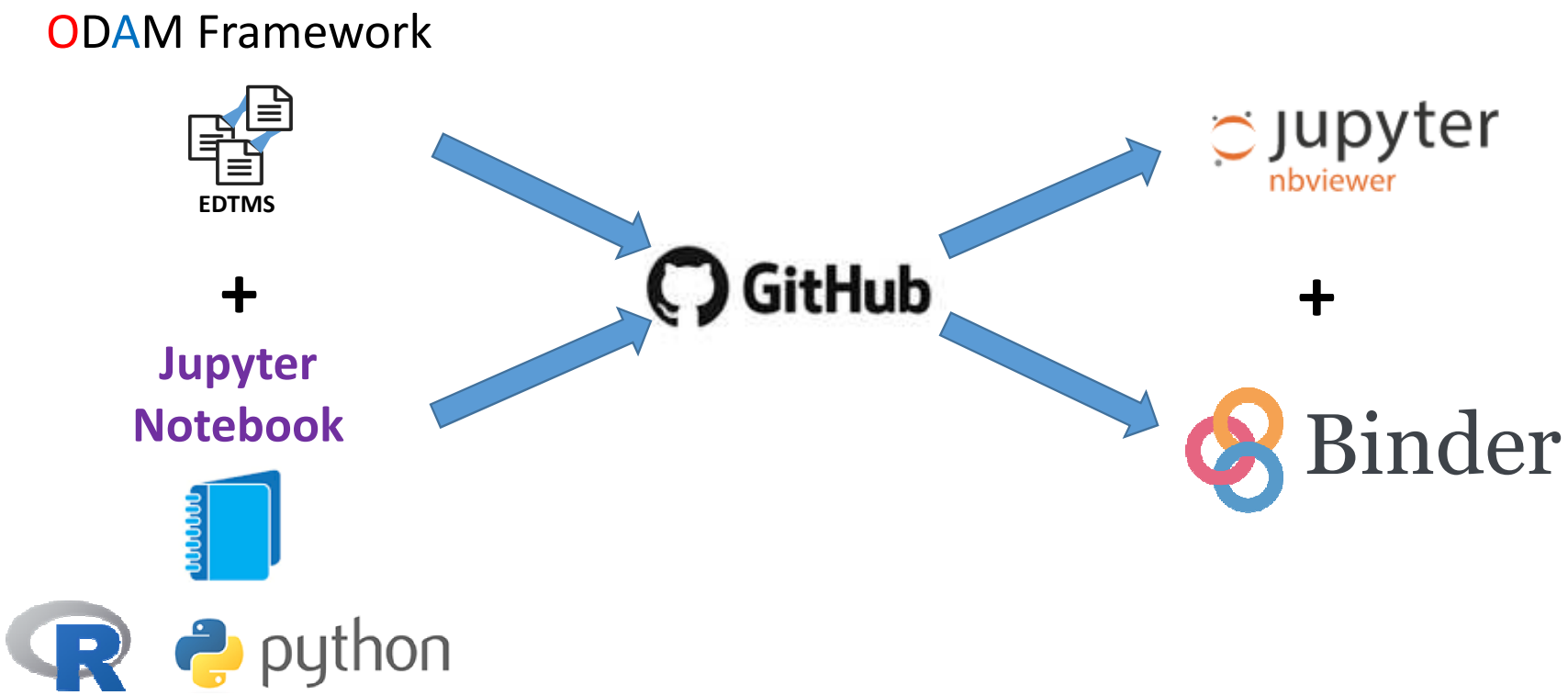
R packages for data science https://www.tidyverse.org/

In particular dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges

INTELLIGENCY http://intelligency.org/ai_r.php

Tools for data science

Reproducible Research



Jupyter Notebook

jupyter nbviewer

JUPYTER FAQ

binder_odam / Rodam_api_graphics.ipynb

In [8]:
```
# See https://www.datacamp.com/community/blog/jupyter-notebook-r
library(Rodam)
library(UpSetR)
library(data.tree)
library(dendextend)
```

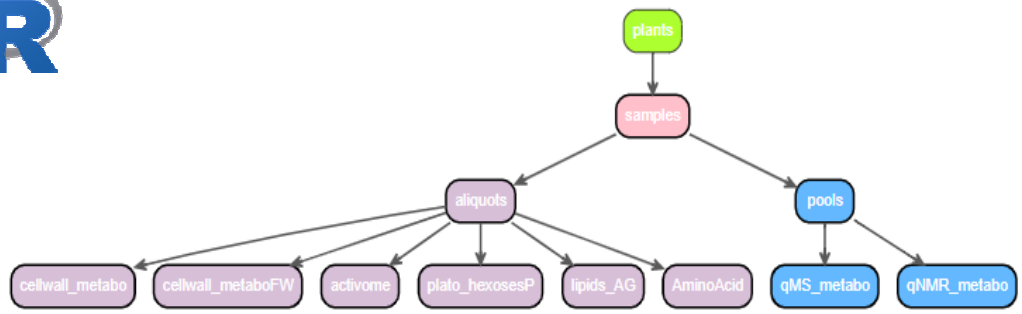
Initialize the 'ODAM' object

In [9]:
```
dh <- new('odamws', wsURL='https://pmb-bordeaux.fr/getdata/', dsname='frim1')
```

ODAM EDTMS

Plot the data tree of the whole dataset

In [11]:
```
dn <- dh$getDataTree()
data.tree::SetGraphStyle(dn, rankdir = "TB")
data.tree::SetEdgeStyle(dn, arrowhead = "vee", color = "grey35", penwidth = 2)
data.tree::SetNodeStyle(dn, style = "filled,rounded", shape = "box", fillcolor = "GreenYellow",  fontname = "helvetica", tooltip = (
data.tree::SetNodeStyle(dn$samples, fillcolor = "Pink", penwidth = "2px")
data.tree::SetNodeStyle(dn$samples$aliquots, fillcolor = "Thistle", penwidth = "2px")
data.tree::SetNodeStyle(dn$samples$pools, fillcolor = "SteelBlue1", penwidth = "2px")

plot(dn)
```



Get the data from ODAM

In [3]:
```
## API Call to retrieve data
dataset = 'frim1'
subset = 'qNMR_metabo'
d = getSubsetFromODAM(dataset, subset)

## Retrieve factors
d['factor']['Attribute']
```

Out[3]:
```
0    Treatment
1     DevStage
2     FruitAge
Name: Attribute, dtype: object
```

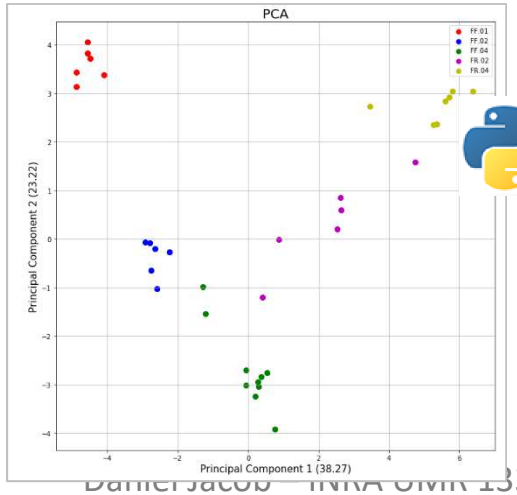
ODAM EDTMS

In [4]:
```
# Matrix X
X = d['data'][d['numvars']]

# Choose 'DevStage' as Factor (index=1)
facname = d['factor'].Attribute[1]
Y = d['data'][facname]

# Factor levels
factorlevels = []
for f in Y:
    if f not in factorlevels:
        factorlevels.append(f)
factorlevels
```

Out[4]:
```
['FF.01', 'FF.03', 'FR.01', 'FR.04', 'FR.02', 'FF.04', 'FF.02', 'FR.03']
```

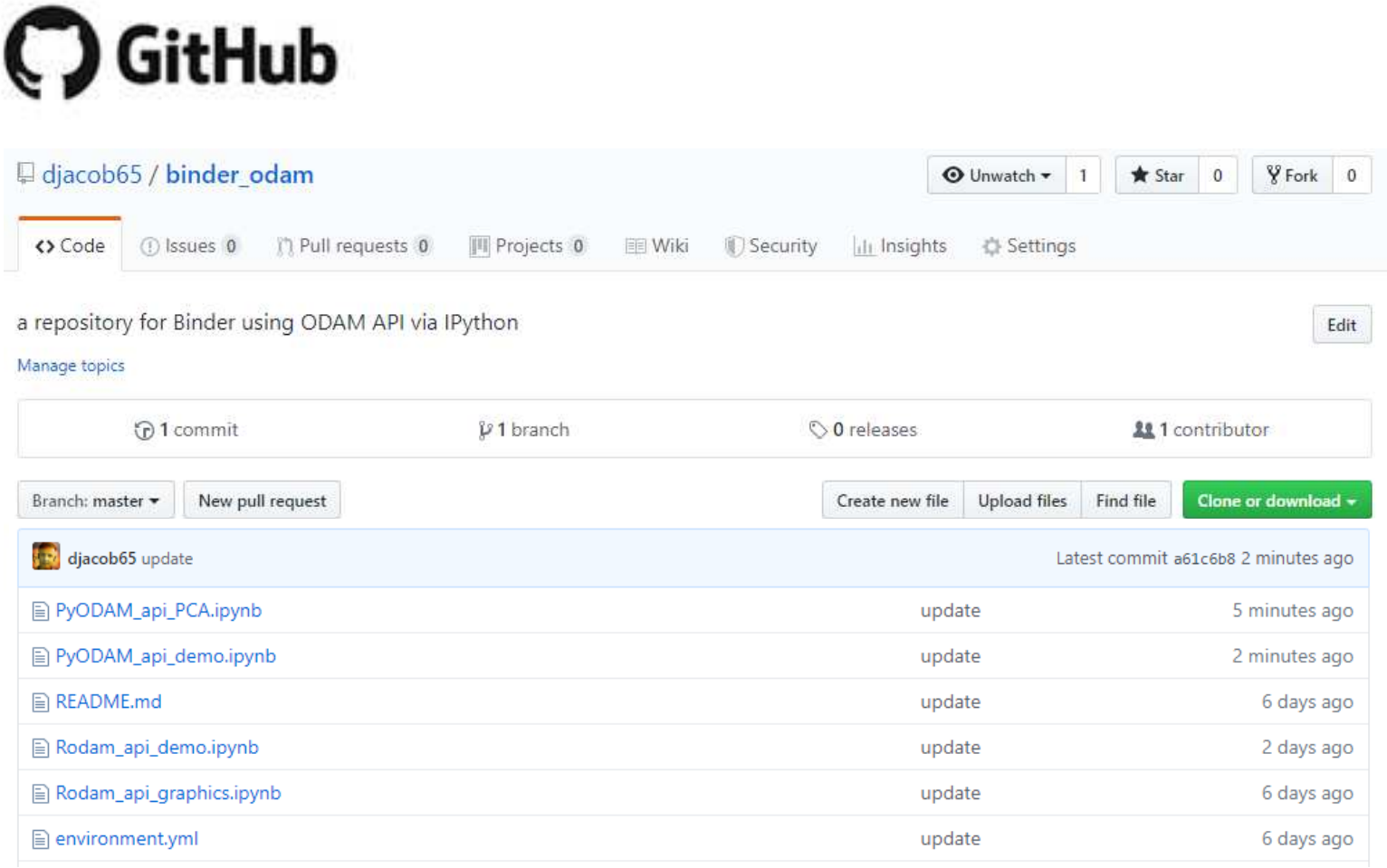
Compute then plot PCA

In [5]:
```
# Compute PCA
res_pca = PCA_compute(X, Y, n=3, scale=True)

# Factor level selection
FacLevSel = ['FF.01', 'FF.02', 'FF.04', 'FR.02', 'FR.04']

# Plot PCA Scores
plotPCA(res_pca, 1, 3, FacLevSel)
```



python

Reproducible Research Jupyter Notebook



environment.yml

```
name: ipython-environment
channels:
  - conda-forge
dependencies:
  - requests
  - numpy
  - matplotlib
  - pandas
  - scikit-learn
```

https://github.com/djacob65/binder_odam

Notebook Viewer https://nbviewer.jupyter.org/

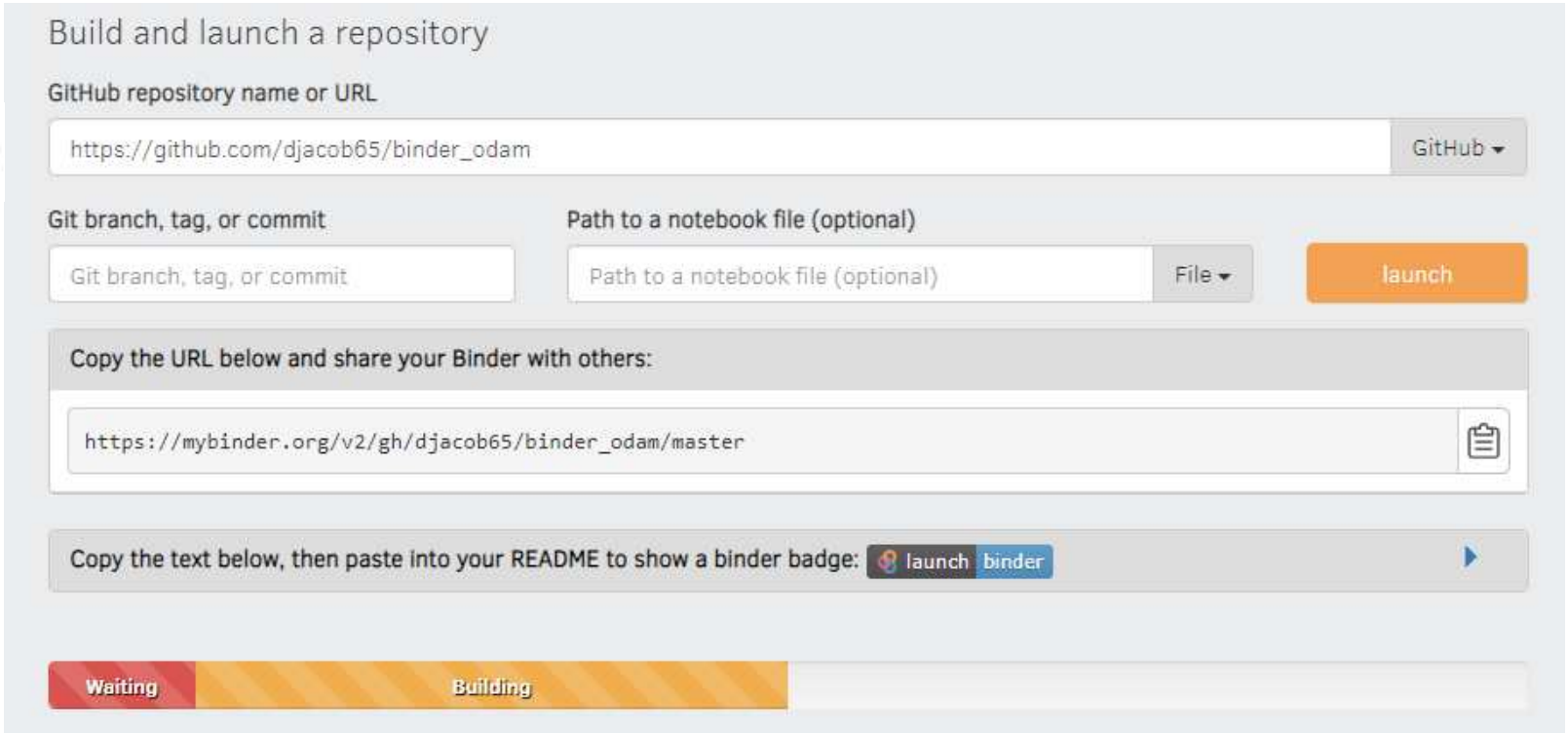
https://nbviewer.jupyter.org/github/djacob65/binder_odam/

Binder

The Binder Project is an open community that makes it possible to create sharable, interactive, reproducible environments.

https://mybinder.org/

GitHub

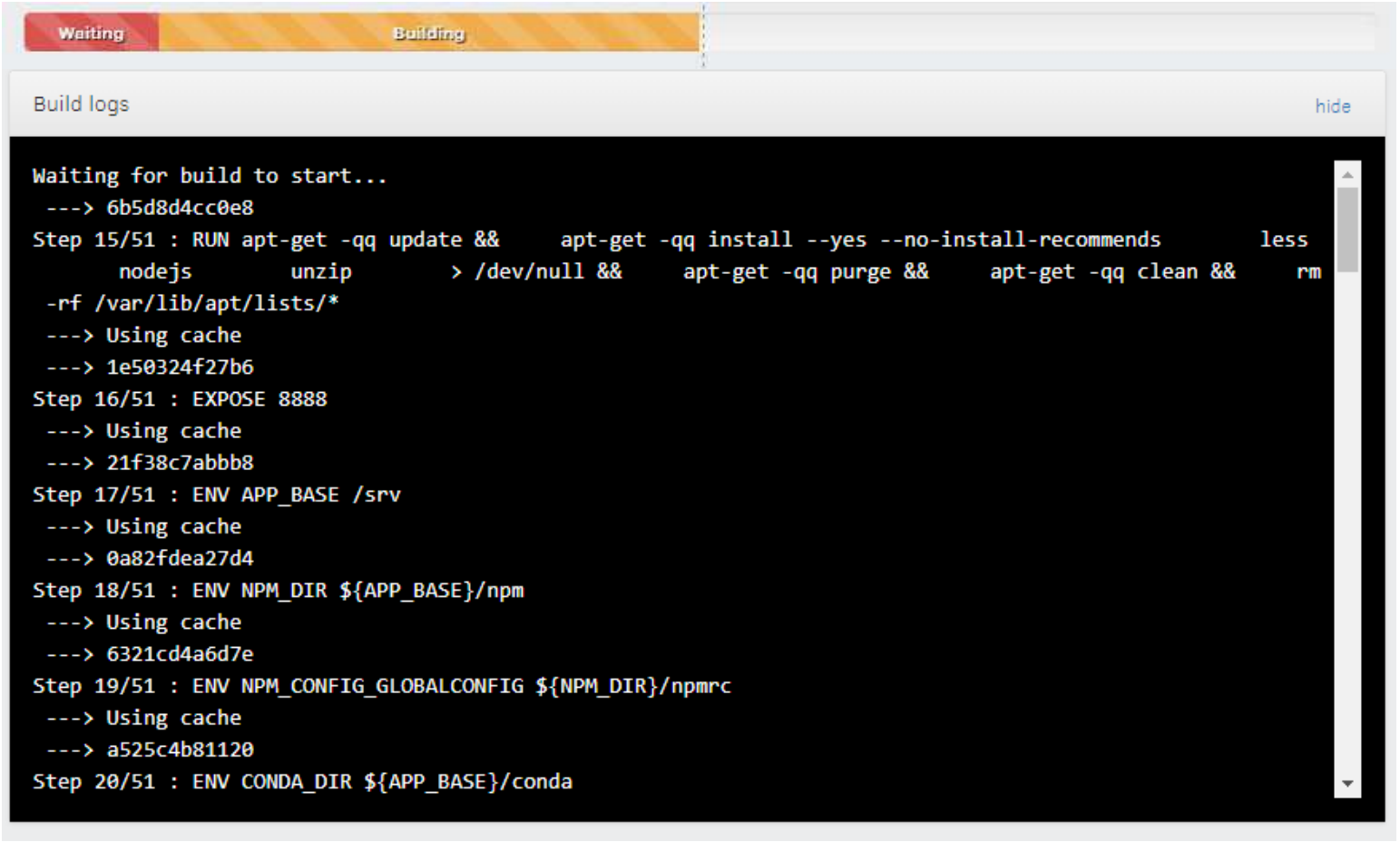


https://mybinder.readthedocs.io/en/latest/introduction.html

Binder

The Binder Project is an open community that makes it possible to create sharable, interactive, reproducible environments.

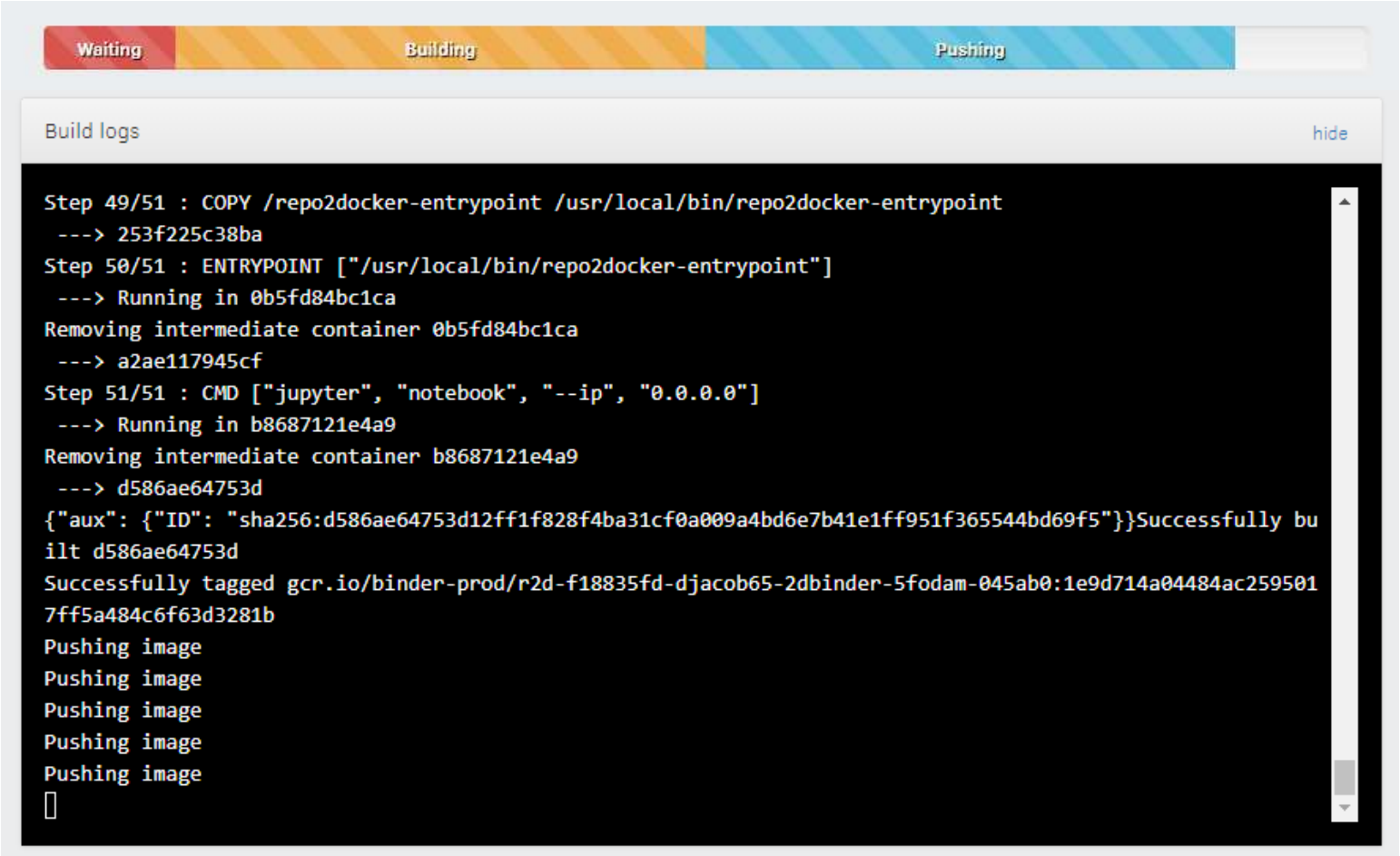
Build the docker container



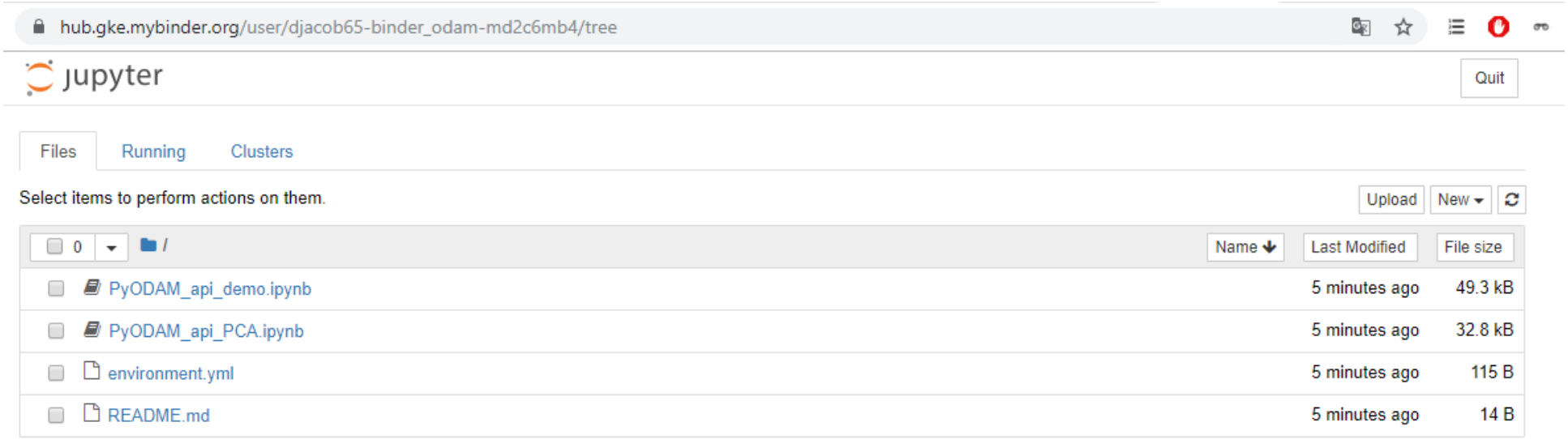
Binder

The Binder Project is an open community that makes it possible to create sharable, interactive, reproducible environments.

Push the docker container into the several nodes



Binder Jupyter Interactive Notebook



Shareable link launch binder

https://mybinder.org/v2/gh/djacob65/binder_odam/master

See more complete examples :

https://github.com/nuest/reproducible-research-and-giscience

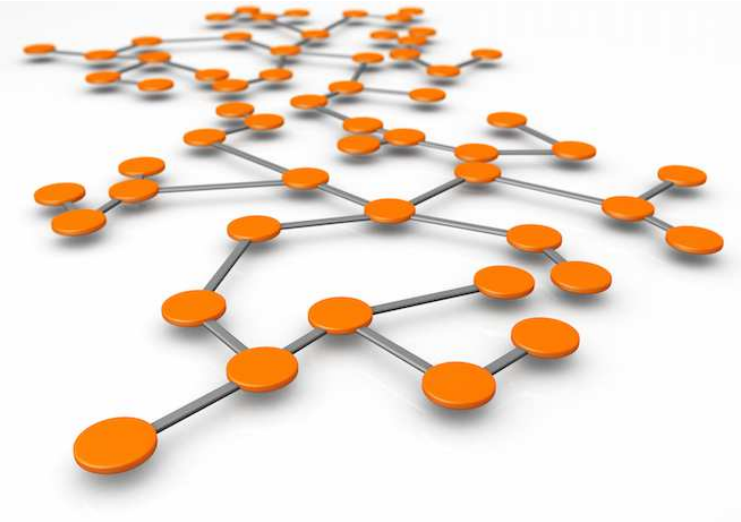
https://ajstewartlang.github.io/SIPS_2019/SIPS_presentation.html

ODAM Framework
Open Data for Access and Mining

https://fr.slideshare.net

"Make your data great now" towards **Open Data** and **Reproducible Research**

"Make your data great again" towards **Linked Open Data**



https://bio.tools/ODAM

Thank you for your attention