How to build the metadata files

ODAM

EDTMS

Tutorial

Daniel Jacob
UMR 1332 BFP – Metabolism Group
Bordeaux Metabolomics Facility
May 2016

Explanation based on an example

FRIM - Fruit Integrative Modelling

See http://www.erasysbio.net/index.php?index=266 & http://frim.brookes.ac.uk/Home

Design of Experiment (DoE)

- 1 Genotype MoneyMaker, Species Lyco. Sola.
- ~600 plants, >1500 samples
- Tissues : Fruit (F)
- Treatment : Control (TC), Shadow (TS), Water Shortage (TWS)
- bouquet, weight, height, diameter,
- Dev. Stages, 9 levels: from march up to august
- Compounds: 12 quantified metabolites
- Enzymes: 38 quantified enzymes

- 5 object types of study, namely: plants, harvests, samples, compounds and enzymes
- 2 factors: Treatment, Development stages
- 53 quantitative variables: compounds (12) + enzymes (38) + weight, height, diameter(3)

FRIM - Fruit Integrative Modelling

seeding

harvesting

samples preparation

Sample identifiers

samples analysis

Compounds

Plants

Harvests

Samples

Enzymes

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	310	E34	E	Control
Money Maker_WT	314	E38		
Money Maker_WT	374	H29		
Money Maker_WT	379	H34		
Money Maker_WT	397	H52		
Money Maker_WT	406	H61		
Money Maker_WT	415	F1		
Money Maker_WT	438	F24		
Money Maker_WT	460	F46		
Money Maker_WT	486	G3		
Money Maker_WT	494	G11		
Money Maker_WT	511	G28		
Money Maker_WT	524	G41		
Money Maker_WT	548	G65		
Money Maker_WT	28	A28		

harvests.tsv

PlanteID	Lot	Truss	FruitAge	HarvestDate	HarvestHour	FruitPositior	FruitFW
18	5	Truss_5	00.08DPA	40379	0.51875	4	1.5
155	5	Truss_5	00.08DPA	40379	0.51875	4	1.0
164	5	Truss_5	00.08DPA	40379	0.51875	4	1.0
221	5	Truss_5	00.08DPA	40379	0.51875	2	1.0
226	5	Truss_5	00.08DPA	40379	0.51875	3	0.8
322	5	Truss_5	00.08DPA	40379	0.51875	2	0.7
343	5	Truss_5	00.08DPA	40379	0.51875	4	0.6
361	5	Truss_5	00.08DPA	40379	0.51875	5	1.0
369	5	Truss_5	00.08DPA	40379	0.51875	4	0.7
372	5	Truss_5	00.08DPA	40379	0.51875	5	1.2
448	6	Truss_5	00.08DPA	40379	0.52430556	5	1.0

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

compounds.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

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						1830.77	270.3	1008.07	778.47	1271.77	965.12	530.51	68.32	9
115						719.39	250.87	1223.19	360.15	526.31	437.56	548.96	194.09	84
117						1168.66	246.51	1039.7	526.42	606.25	607.02	557.28	163.25	1
119						1035.6	260.67	1019.02	663.14	986.63	632.42	579.25	NA	90
361						1245.73	243.07	915.78	530.86	832.47	659.72	518.63	91.02	87
						363.23	131.41	525.24	268.55	337.81	210.84	581.39	118.38	61
						450.13	171.89	593.3	357.74	220.39	247.36	587.58	20.21	86
						330.26	131.87	499.38	208.85	249.95	190.41	545.06	58.85	57
						573.06	128.65	390.78	283.25	304.36	301.7	NA	168.46	83
						286.02	120.11	438.09	260.44	240.08	204.19	500.82	88.22	
19	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
20	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
21	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
22	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

Data subsets files must be compliant with the **TSV standard** (Tab-Separator-Values)

INRA SCIENCE & IMPACT

FRIM - *Preparation and cleaning of the data sub-sets of files*

Data subset files

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	310	E34	E	Control

harvests.tsv

PlanteID	Lot	Truss	FruitAge	HarvestDate	HarvestHour	FruitPositior	FruitFW	FruitDiamete	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

enzymes.tsv

ISODH_NAD	EnoS	ISODH_NADI	PEPC
508.99	NA	NA	93
975.77	514.4	133.6	11
555.54	588.57	167.86	99
572.33	676.61	87.43	27
421.23	504.23	231.52	81
416.04	530.9	59.01	8
660.68	645.49	221.32	68
965.12	530.51	68.32	9
437.56	548.96	194.09	84
607.02	557.28	163.25	1
632.42	579.25	NA	90
659.72	518.63	91.02	87
210.84	581.39	118.38	61
247.36	587.58	20.21	86
190.41	545.06	58.85	57
301.7	NA	168.46	83
204.19	500.82	88.22	
366.29	639.19	45.01	127
256.53	766.52	NA	6
202.93	600.66	41.36	9
146.66	660.62	124.38	58

- 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 experimental factors.
- Moreover, the objects of study have usually an identifier for each of them, and the variables can be quantitative or qualitative.
- We can have either one object type of study or several kinds, but in this latter case, it must exist a relationship between object types that we assume of "**obtainedFrom**" type.

FRIM - *Classification of each column within its right category*

Data subset files

plants.tsv

Genotype	PlanteID	Plant	Row	Treatment
Money Maker_WT	26	A26	A	Control
Money Mak	140	C2	C	Control
Mone _WT	222	D15	D	Control

harvests.tsv

	A	B	C	D	E	F	G	H	I	J
1	PlanteID	Lot	Truss	FruitAge	HarvestDate	HarvestHour	FruitPositior	FruitFW	FruitDiamet	FruitHeight
30	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
32	164	5	Truss_5	00.08DPA						
33	221	5	Truss_5	00.08DPA						
34	226	5	Truss_5	00.08DPA						
35	322	5	Truss_5	00.08DPA						

samples.tsv

	A	B	C	D	E	F
1	Lot	SampleID	NbFruit	GellyFW	GellyFruit	BER
2	1	1	10	2.83	0.283	FALSE
3	1	2	10	2.83	0.283	FALSE

compounds.tsv

	A		C	D	E	F	G	H	I	J	K	L	M	N	O
1	SampleID	DPA	MassBefore	MassMIA	RDT	Starch1	Starch2	RHAMNOSE	ARABINOSE	XYLOSE	MANNOSE	GALACTOSE	GLUCOSE	OsesN	PoidsAU
2	363	15	0.192		52.19	36.55	0.19074531	0.89	2.17	1.73	1.2	5.99	11.31	23.29	7.17
3	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
4	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
5	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
6	429	28	0.104	0.0431	41.44	19.78	0.08197289	0.79	2.13	2.33	2.42	8.51	15.2	31.38	9.94
7	431														
8	433														
9	435														
10	445														
11	451														
24	1169														
25	1187														
26	1343														
27	1345														
28	1347														
29	1349														
30	1351														
31	1353														
32	1355	8	0.093	0.057	61.2903226	29.63	0.18160323	0.64	2.25	2.57	3.2	3.65	15.73	28.04	

enzymes.tsv

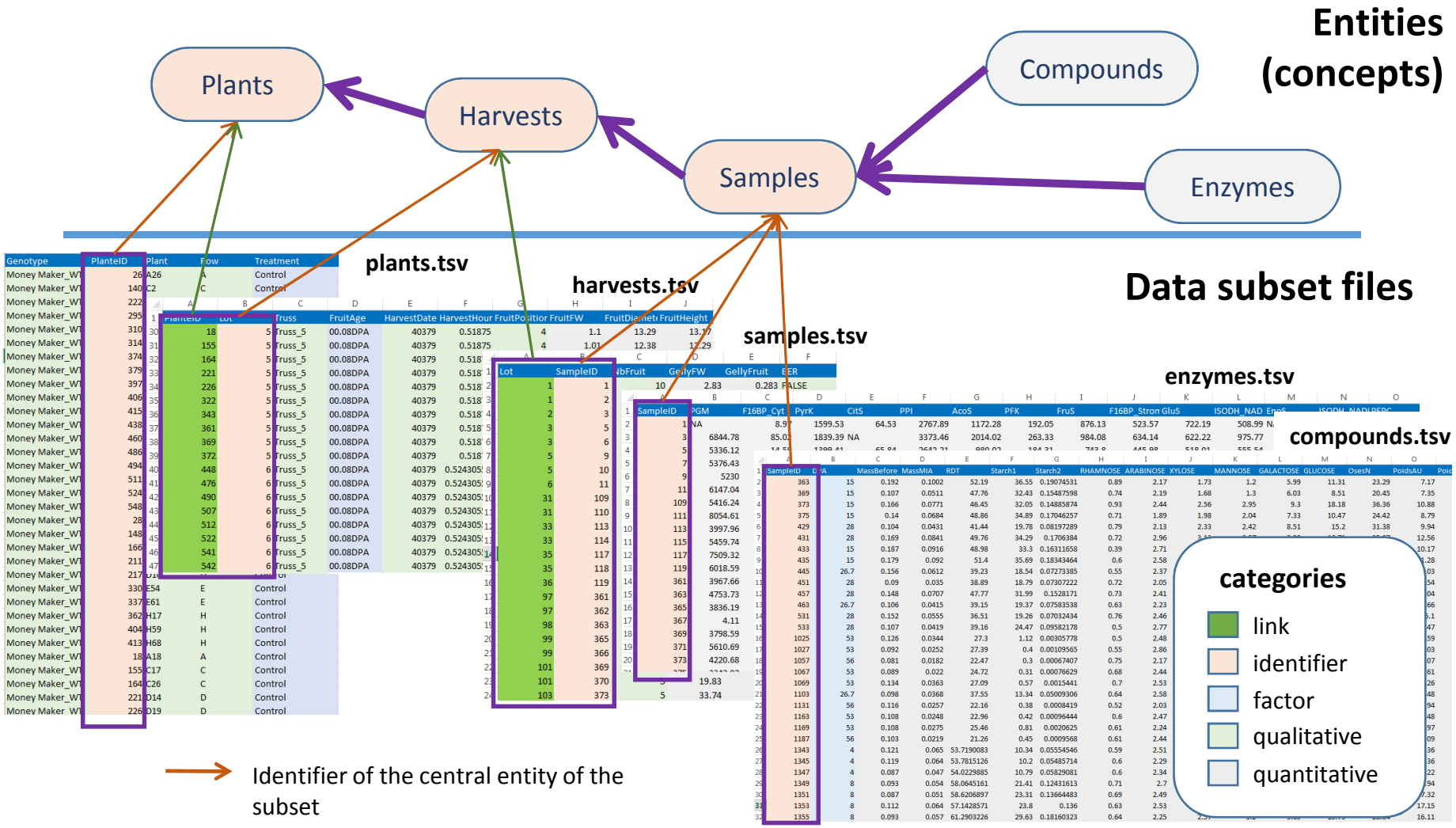
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	SampleID	PGM	F16BP_Cyt	PyrK	CitS	PPI	AcoS	PFK	FruS	F16BP_Stron	GluS	ISODH_NAD	EnoS	ISODH_NADI	PEPC
2	1	NA	8.97	1599.53	64.53	2767.89	1172.28	192.05	876.13	523.57	722.19	508.99	NA	NA	936.41
3	3	6844.78	85.02	1839.39	NA	3373.46	2014.02	263.33	984.08	634.14	622.22	975.77	514.4	133.6	1107.8
15	363	4753.73	17.77	728.5	NA	1869.46	450.13	171.89	593.3	357.74	518.01	555.54	588.57	167.86	997.01
16	365	3836.19	4.16	603.74	31.55	1836.44	330.26	131.87	499.38	208.85	915.19	572.33	676.61	87.43	273.91
17	367	4.11	4.87	507.67	7.39	1371.13	573.06	128.65	390.78	283.25	510.8	421.23	504.23	231.52	813.03
18	369	3798.59	15.14	703.9	82.78	1473.25	286.02	120.11	438.09	260.44	598.68	416.04	530.9	59.01	855.6
19	371	5610.69	14.69	1092.42	99.09	2384.35	799.55	182.18	812.09	425.56				221.32	686.05
20	373	4220.68	31.7	592.57	14.81	1654.27	602.62	150.23	480.9	283.92					994.7
															845.34
															130.2
															908.46
															871.25
															615.66
															864.75
															571.97
															830.04
															763
															1271.88
															695.6

You have to organize your data subsets so that links could be established between them. In practical, it means to add a column containing the identifiers corresponding to the entity to which you want to connect the subset. It is to be noted that this duplication of identifiers must be the only redundant information, through all data subsets.

categories

- link
- identifier
- factor
- qualitative
- quantitative

FRIM - *Connections between the dataset files based on identifiers*



FRIM - *Creation of the metadata files*

Supplementary files

In order to allow data to be explored and mined, we have to adjoin some minimal but relevant metadata:

For that, **2 metadata files** are required

- **s_subsets.tsv**: a file allowing to associate with each subset of data a key concept corresponding to the main entity of the subset and the relations of the type "obtainedFrom" between these concepts

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

Note: Data subsets files and their associated metadata files must be compliant with the **TSV standard** (Tab-Separator-Values)

TSV is an alternative to the common comma-separated values (CSV) format, which often causes difficulties because of the need to escape commas

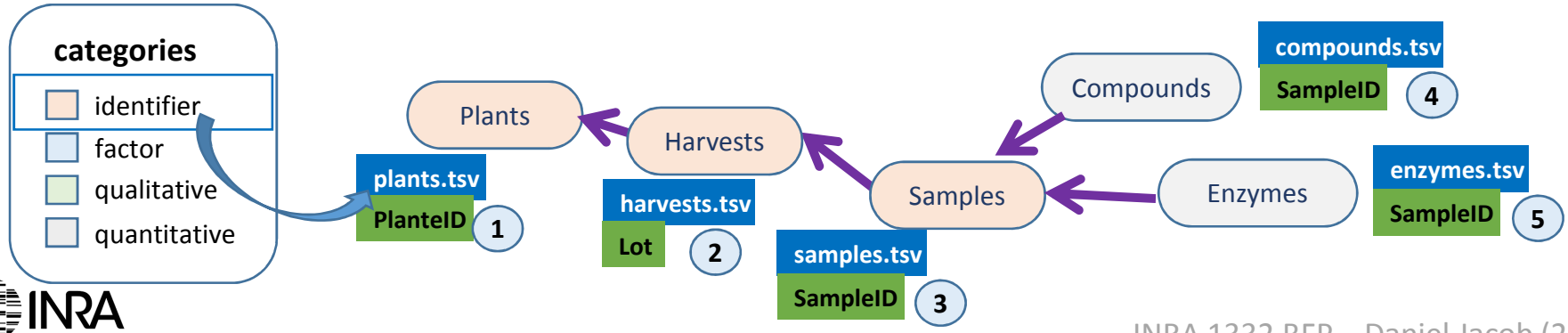
FRIM - *Creation of the metadata files*

s_subsets.tsv This metadata file allows 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.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

1 → Plants | PlanteID | plants.tsv

- plants.tsv: Data file name
- PlanteID: Identifier of the central entity of the subset. This identifier must be present as a column name in the corresponding data subset (e.g. plants.tsv) and also as an attribute within the a_attributes.tsv file in the 'attribute' column (see further).
- Plants: Key concept (i.e. the main entity) associated to the subset in the form of a short name
- obtainedFrom: Link between 2 subsets (implies a '**obtainedFrom**' relation)
- rank: Unique rank number of the data subset



FRIM - *Creation of the metadata files*

s_subsets.tsv This metadata file allows 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.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

a description for each subset

Optional:
an annotation based on ontology

FRIM - *Creation of the metadata files :* Be careful in the spelling

s_subsets.tsv This metadata file allows 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

1. For short names of the data subsets (**C column**), the name of the identifier attributes (**D column**) and the names of the files (**E column**),
   - only the alphanumerical characters and the underscore are allowed (i.e. 'a-z', 'A-Z', '0-9' and '_').
   - Moreover, these names should not start with a digit!
2. For description (**F column**),
   - the allowed characters are : 0-9 a-z A-Z , : + * () [] {} - % ! | / . ?

The identifier attributes (**D column**)

1. should be the only attribute declared as '***identifier***' in the '**category**' column in the a_attributes.tsv file (**D column**)
2. should be available as a column item in the corresponding data subset file

FRIM - *Creation of the metadata files*

a_attributes.tsv This metadata file allows each attribute (variable) to be annotated with some minimal but relevant metadata

Plants

Harvests

Samples

Compounds

...

	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		Data Type
5	plants	Treatment	treatment	factor	string	Treatment appl		
6	plants	Genotype		qualitative	string	Genotype		
7	harvests	Lot	lot	identifier	numeric	Pool of severa		
8	harvests	PlanteID			numeric	Plant identifie		r
9	harvests	Truss		qualitative	string	Position on th		
10	harvests	HarvestDate			string	Harvest date		
11	harvests	HarvestHour			string	Harvest hour		
12	harvests	FruitAge	age	factor	string	Fruit developn		
13	harvests	FruitPosition		qualitative	numeric	Poistion on th		
14	harvests	FruitDiameter		quantitative	numeric	Fruit diameter		
15	harvests	FruitHeight		quantitative	numeric	Fruit height (n		
16	harvests	FruitFW		quantitative	numeric	Fruit Fresh We		
17	samples	SampleID	sampleid	identifier	numeric	Sample identif		er
18	samples	Lot			numeric	Pool of several ha		
19	samples	NbFruit			numeric	Fruit Number per sample		
20	samples	GellyFW		quantitative	numeric	le		
21	samples	GellyFruit			numeric			
22	samples	BER			string			
23	compounds	SampleID			numeric		http://purl.obolibrary.org	centrally registered identifier
24	compounds	DPA			numeric			
25	compounds	MassBefore			numeric			
26	compounds	MassMIA			numeric			
27	compounds	RDT			numeric			
28	compounds	Starch1			numeric		http://purl.obolibrary.org	starch
29	compounds	Starch2			numeric		http://purl.obolibrary.org	starch
30	compounds	RHAMNOSE			numeric		http://purl.obolibrary.org	rhamnose

1. Short names of the data subsets (**A column**) must be declared in the s_subsets.tsv file (**C column**) and vice versa.
2. One and only one attribute (**B column**) must be declared as '***identifier***' in the '**category**' column (**D column**) per data subset (**A column**)

the attribute names

the attribute data types

...

FRIM - *Creation of the metadata files*

a_attributes.tsv This metadata file allows each attribute (variable) to be annotated with some minimal but relevant metadata

plants.tsv

Plants

Harvests

Samples

Compounds

...

	A	B	C	
1	subset	attribute	entry	cate
2	plants	PlanteID	plantid	ider
3	plants	Row	row	qual
4	plants	Plant	plant	
5	plants	Treatment	treatment	fact
6	plants	Genotype		qual
7	harvests	Lot	lot	ider
8	harvests	Plant ID		
9	harvests	Truss		qual
10	harvests	HarvestDate		
11	harvests	HarvestHour		
12	harvests	FruitAge	age	fact
13	harvests	Fru		qual
14	harvests	Fru		qual
15	harvests	Fru		qual
16	harvests	FruitFW		qual
17	samples	SampleID	sampleid	ider
18	samples	Lot		
19	samples	NbFruit		
20	samples	GellyFW		qual
21	samples	GellyFruit		qual
22	samples	BER		
23	compounds	SampleID	sampleid	ider
24	compounds	DPA		fact
25	compounds	MassBefore		qual
26	compounds	MassMIA		qual
27	compounds	RDT		qual
28	compounds	Starch1		qual
29	compounds	Starch2		qual
30	compounds	RHAMNOSE		qual

Transpose

plants.csv - Excel

FICHIER ACCUEIL INSERTION MISE EN PAGE FORMULES DONNÉES RÉVIS

A1 Genotype

	A	B	C	D	E	F
1	Genotype	PlanteID	Plant	Row	Treatment	
2	Money Make	26	A26	A	Control	
3	Money Make	140	C2	C	Control	
4	Money Make	222	D15	D	Control	
5	Money Make	295	E19	E		
6	Money Make	310	E34	E		
7	Money Make	314	E38	E		
8	Money Make	374	H29	H		
9	Money Make	379	H34	H		
10	Money Make	397	H52	H		
11	Money Make	406	H61	H	Control	
12	Money Make	415	F1	F	WATER STRESS	
13	Money Make	438	F24	F	WATER STRESS	
14	Money Make	460	F46	F	WATER STRESS	
15	Money Make	486	G3	G	WATER STRESS	
16	Money Make	494	G11	G	WATER STRESS	
17	Money Make	511	G28	G	WATER STRESS	

plants

PRÊT NB (NON VIDES) : 5 100 %

Get attribute names simply from the data subsets

...

FRIM - *Creation of the metadata files*

a_attributes.tsv This metadata file allows each attribute (variable) to be annotated with some minimal but relevant metadata

harvests.tsv

Plants

Harvests

Samples

Compounds

	A	B
1	subset	attribute
2	plants	PlanteID
3	plants	Row
4	plants	Plant
5	plants	Treatment
6	plants	Genotype
7	harvests	Lot
8	harvests	PlanteID
9	harvests	Truss
10	harvests	HarvestDate
11	harvests	HarvestHour
12	harvests	FruitAge
13	harvests	FruitPosition
14	harvests	FruitDiameter
15	harvests	FruitHeight
16	harvests	FruitFW
17	samples	Sample
18	samples	Lot
19	samples	NbFruit
20	samples	GellyFW
21	samples	GellyFruit
22	samples	BER
23	compounds	SampleID
24	compounds	DPA
25	compounds	MassBefore
26	compounds	MassMIA
27	compounds	RDT
28	compounds	Starch1
29	compounds	Starch2
30	compounds	RHAMNOSE

harvests.csv - Excel

FICHIER ACCUEIL INSERTION MISE EN PAGE FORMULES DONNÉES RÉVISION AFFICHAGE PDF Architect Connexion

A1 PlanteID

	A	B	C	D	E	F	G	H	I	J
1	PlanteID	Lot	Truss	FruitAge	HarvestDate	HarvestHour	FruitPositior	FruitFW	FruitDiamet	FruitHeight
30	18	5	Truss_5	00.08DPA	40379	0.51875	4	1.1	13.29	13.17
31	155	5	Truss_5	00.08DPA	40379	0.51875	4	1.01	12.38	12.29
32	164	5	Truss_5	00.08DPA	40379	0.51875	4	1.02	12.28	12.44
33	221	5	Truss_5	00.08DPA	40379	0.51875	2	1.02	12.8	12.28
34	226	5	Truss_5	00.08DPA	40379	0.51875	3	0.81	12.27	11.43
35	322	5	Truss_5	00.08DPA	40379					10.46
36	343	5	Truss_5	00.08DPA	40379					.09
37	[illegible]	5	Truss_5	00.08DPA	40379					.58
38	[illegible]	5	Truss_5	00.08DPA	40379					.72
39	372	5	Truss_5	00.08DPA	40379					.66
40	448	6	Truss_5	00.08DPA	40379	0.[illegible]				.82
41	476	6	Truss_5	00.08DPA	40379	0.52430556	5	0.66	10.99	10.18
42	490	6	Truss_5	00.08DPA	40379	0.52430556	5	0.62	10.72	10.79
43	507	6	Truss_5	00.08DPA	40379	0.52430556	5	0.75	12.14	10.77
44	512	6	Truss_5	00.08DPA	40379	0.52430556	2	0.58	10.92	9.92
45	522	6	Truss_5	00.08DPA	40379	0.52430556	5	0.86	11.88	11.53

harvests

PRÊT 100 %

Transpose

Get attribute names simply from the data subsets

...

...

FRIM - *Creation of the metadata files*

a_attributes.tsv This metadata file allows each attribute (variable) to be annotated with some minimal but relevant metadata

Plants

Harvests

Samples

Compounds

...

	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		st			
5	plants	Treatment	treatment	factor	st			
6	plants	Genotype		qualitative	st			
7	harvests	Lot	lot	identifier	n			
8	harvests	PlanteID			n			
9	harvests	Truss		qualitative	st			
10	harvests	HarvestDate			st			
11	harvests	HarvestHour			st			
12	harvests	FruitAge	age	factor	st			
13	harvests	FruitPosition		qualitative	n			
14	harvests	FruitDiameter		quantitative	n			
15	harvests	FruitHeight		quantitative	n			
16	harvests	FruitFW		quantitative	n			
17	samples	SampleID	sampleid	identifier	n			
18	samples	Lot			numeric	Pool of several harvests	http://www.ebi.ac.uk/ef	sample pooling
19	samples	NbFruit			numeric	Fruit Number per sample		
20	samples	GellyFW		quantitative	numeric	Gelly Fred Weig		
21	samples	GellyFruit		quantitative	numeric	Gelly per Fruit		
22	samples	BER			string	BER		
23	compounds	SampleID	sampleid	identifier	numeric	Sample identi	obolibrary.org	centrally registered identifier
24	compounds	DPA		factor	numeric	Day Per Anthe		
25	compounds	MassBefore		quantitative	numeric	m av.extractio		
26	compounds	MassMIA		quantitative	numeric	masse MIA (g)		
27	compounds	RDT		quantitative	numeric	Rdt (% MIA/D		
28	compounds	Starch1		quantitative	numeric	Dosage amido	obolibrary.org	starch
29	compounds	Starch2		quantitative	numeric	amidon (g/gD	obolibrary.org	starch
30	compounds	RHAMNOSE		quantitative	numeric	RHAMNOSE	obolibrary.org	rhamnose

- Dependent variables resulting of effects of some controlled experimental factors.
- Objects of study have usually an identifier for each of them, and the variables can be quantitative or qualitative.

categories

- identifier
- factor
- qualitative
- quantitative

...

FRIM - *Creation of the metadata files*

a_attributes.tsv This metadata file allows each attribute (variable) to be annotated with some minimal but relevant metadata

Plants

Harvests

Samples

Compounds

	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						
11	harvests	HarvestHour						
12	harvests	FruitAge	age	factor			http://purl.obolibrary.org	fruit development stage
13	harvests	FruitPosition		qualit			http://ncicb.nci.nih.gov/x	Position Number
14	harvests	FruitDiameter		quant			http://ncicb.nci.nih.gov/x	Diameter
15	harvests	FruitHeight		quant			http://ncicb.nci.nih.gov/x	Height
16	harvests	FruitFW		quant			http://ncicb.nci.nih.gov/x	Weight
17	samples	SampleID	sampleid	ident			http://purl.obolibrary.org	centrally registered identifier
18	samples	Lot					http://www.ebi.ac.uk/efo	sample pooling
19	samples	NbFruit						
20	samples	GellyFW		quant				
21	samples	GellyFruit		quant				
22	samples	BER						
23	compounds	SampleID	sampleid	ident			http://purl.obolibrary.org	centrally registered identifier
24	compounds	DPA		factor				
25	compounds	MassBefore		quant				
26	compounds	MassMIA		quant				
27	compounds	RDT		quant				
28	compounds	Starch1		quant			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

Give opportunity to make a selection on the attributes via web-services by associating them an alias name (called an "entry")

...

http://myhost.org/getdata/xml/frim1/(samples)/treatment/Control

FRIM - *Creation of the metadata files*

a_attributes.tsv This metadata file allows each attribute (variable) to be annotated with some minimal but relevant metadata

Plants

Harvests

Samples

Compounds

...

	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		q				
22	samples	BER						
23	compounds	SampleID	sampleid	id			h	
24	compounds	DPA		fa				
25	compounds	MassBefore		q				
26	compounds	MassMIA		q				
27	compounds	RDT		q				
28	compounds	Starch1		q		A)	h	
29	compounds	Starch2		q			h	
30	compounds	RHAMNOSE		q			h	

a description for each attribute

Optional: an annotation based on ontology

FRIM - *Creation of the metadata files :* Be careful in the spelling

a_attributes.tsv This metadata file allows each attribute (variable) to be annotated with some minimal but relevant metadata

	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	ha							
10								

1. For short names of the data subsets (**A column**), the name of the attributes (**B column**) and the entry names (**C column**):
   - only the alphanumerical characters and the underscore are allowed (i.e. 'a-z', 'A-Z', '0-9' and '_').
   - Moreover, these names should not start with a digit!
2. For categorical names (**D column**):
   - the set of terms are fixed, namely: '***identifier***', '***factor***', '***quantitative***', '***qualitative***'. Leave as blank otherwise.
3. For data types (**E column**):
   - the allowed names are restricted to '***numeric***' or '***string***'.
4. For description (**F column**):
   - the allowed characters are : 0-9 a-z A-Z , : + * () [] {} - % ! | / . ?

FRIM - *Updating the metadata files*

s_subsets.tsv

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

...

a_attributes.tsv

subset	attribute	entry	category	type	description	
plants	PlanteID	plantid	identifier	numeric	Plant identifier	
plants	Row	row	qualitative	string	Row of the invidual plant or	
plants	Plant	plant		string	Code identifier of the indiv	
plants	Treatment	treatment	factor	string	Treatment applied on plant	
plants	Genotype		qualitative	string	Genotype	
harvests	Lot	lot	identifier	numeric	Pool of several harvests	
harvests	PlanteID			numeric	Plant 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

...

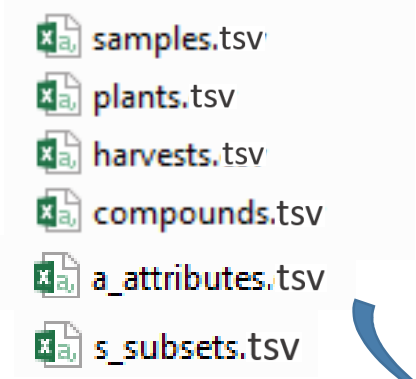
Additional subsets/ attributes can be added step by step, as soon as data are produced.

FRIM - *Uploading your datasets in the data repository*

Merely dropping data files on the data repository (e.g. NAS) should allow users to access them by web services

Data capture

Minimal effort (PUT)

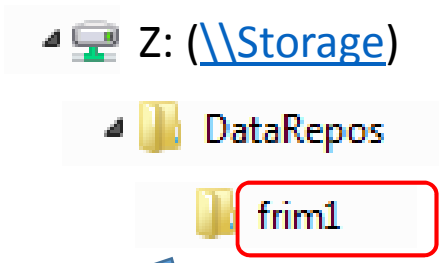


Your data subset files

PUT

Data subsets files and their associated metadata files must be compliant with the **TSV standard** (Tab-Separator-Values)

Data repository



Your dataset entry (named 'frim1' as example) within the data repository

mount

myhost.org

http://myhost.org/

GET

Data analysis/mining

Maximum efficiency (GET)

No database schema, no programming code and no additional configuration on the server side.

FRIM - *Checking online if your the data subset files are consistent*

Before dropping your data set in the data repository, check all points below:

Ref. Note	Note Description
1	A directory named as the dataset name should be actually created in the data repository; Be careful in the spelling, see note 6;
2	The s_subsets.tsv and a_attributes.tsv files should be present in the data repository
3	All data subset files declared in the s_subsets.tsv (col 5) should be available in the data repository
5	To be sure to have the right format, do a 'copy' of data from the spreadsheet then 'paste' them into a new file, then 'save as TSV format (separator: a tab character)'
4	1) all subsets in the a_attributes.tsv file (col 1) should be declared in the s_subsets.tsv file (col 3) 2) all subsets in the s_subsets.tsv file (col 3) should be declared in the a_attributes.tsv file (col 1) 3) all attribute names in the a_attributes.tsv file (col 2) should be available as a column in the corresponding data subset file declared in the s_subsets.tsv file (col 5)
6	Be careful in the spelling: 1) for data subset file names (col 5 in s_subsets.tsv), identifier name (col 2 in s_subsets.tsv), attribute names (col 2 in a_attributes.tsv), subset shortnames (col 3 in s_subsets.tsv and col 1 in a_attributes.tsv) and entry names (col 3 in a_attributes.tsv), only the alphanumerical characters and the underscore are allowed (i.e 'a-z', 'A-Z', '0-9' and '_'). Moreover, these names should not start with a digit! 2) for categorical names (col 4 in a_attributes.tsv), the number of terms and their spelling are fixed, namely: 'identifier','factor','quantitative','qualitative'. 3) for type (col 5 in a_attributes.tsv), the allowed names are restrited to 'numeric' or 'string'. 4) for descritpion, the allowed characters are : 0-9 a-z A-Z , : + * () [] {} - % ! | / . ?
7	Identifiers declared in the s_subsets.tsv file (col 2) 1) should be declared as 'identifier' in the 'category' column in the a_attributes.tsv file (col 4) 2) should be available as an column item in the corresponding data subset file 3) should be the only one attribute declared as identifier for the corresponding data subset file in the a_attributes.tsv file (col 4)
8	Each subset having a 'father_rank' greater than 0 in the s_subsets.tsv file (col 2) 1) should have a column label in the corresponding data file that must be identical to the identifier label of the linked subset (i.e. corresponding to the 'father_rank' in col 1) 2) should have the linked subset identifier with no category (i.e. void) in the a-attributes.tsv file (col 4), except if the subset and the linked subset have the same identifier

FRIM - *Checking online if your the data subset files are consistent*

Fortunately, most of checking can be automatically done for you

Checklist http://myhost.org/check/frim1

Type	Description	Information	Status
General	the dataset directory	Available	ok
General	the subset definition file	Available	ok
General	the attribute definition file	Available	ok
General	data subset files	all availables (11)	ok
s_subsets	definition file format	CSV-compliant	ok
a_attributes	definition file format	CSV-compliant	ok
a_attributes	Check if spelling of names are proper in a_attributes.tsv	all attributes seem ok	ok
plants.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
samples.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
aliquots.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
compounds.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
enzymes.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
pools.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
FRIM1Quantities.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
qnmr_metabo.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
plato_HexosesP.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
lipids_ag.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
aminoacids.txt	Check if column names are the same that those declared in a_attributes.tsv	all attributes seem ok	ok
identifiers	Check if identifiers are consistent	all identifiers seem consistent	ok
subsets	Check SQL on each subset	all SQL success	ok
merged_subsets	Check SQL on each merged subset	all SQL success	ok

FRIM - *Testing online if your the data subset files are OK*

http://myhost.org/xml/frim1

Subset	Description	Identifier	WSEntry	SetID	LinkID	CV_Term_ID	CV_Term_Name
plants	Plant features	PlantID	plant	1	0	http://purl.obolibrary.org/obo/PO_0000003	whole plant
samples	Sample features	SampleID	sample	2	1	http://purl.obolibrary.org/obo/OBI_1110046	organ harvesting
aliquots	Aliquots features	AliquotID	aliquot	3	2	http://purl.obolibrary.org/obo/PO_0009001	fruit
cellwall_metabo	Cell wall Compound quantifications	AliquotID	aliquot	4	3	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
cellwall_metaboFW	Cell Wall Compound quantifications (FW)	AliquotID	aliquot	5	3	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
activome	Activome Features	AliquotID	aliquot	6	3	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
pools	Pools of remaining pools	PoolID	pool	7	2	http://purl.obolibrary.org/obo/OBI_1110046	organ harvesting
qMS_metabo	MS Components quantification	PoolID	pool	8	7	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
qNMR_metabo	Pools of remaining pools	PoolID	pool	9	7	http://purl.obolibrary.org/obo/OBI_1110046	organ harvesting
plato_hexosesP	Hexoses Phosphate	AliquotID	aliquot	10	3	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
lipids_AG	Lipids AG	AliquotID	aliquot	11	3	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity
AminoAcid	Amino Acids	AliquotID	aliquot	12	3	http://purl.obolibrary.org/obo/CHEBI_24431	chemical entity

To summarize

1. *Preparation and cleaning of the data sub-sets of files*
2. *Classification of each column within its right category*
3. *Connections between the dataset files based on identifiers*
4. *Creation of the metadata files:* ***s_subsets.tsv*** *and* ***a_attributes.tsv***
5. *Deposit of the dataset files in the data repository*
6. *Checking online if your the data subset files are consistent*
7. *Testing online the web-services on your dataset*

Note: Data subsets files and their associated metadata files must be compliant with the **TSV standard** (Tab-Separator-Values)

TSV is an alternative to the common comma-separated values (CSV) format, which often causes difficulties because of the need to escape commas