Make your data great now & again

Comment gérer au mieux nos données pour en tirer le meilleur parti pour nos recherches ?

Faire en sorte que les données ouvertes fonctionnent pour la recherche et les chercheurs ?

Question préliminaire :

Capture des données ?

Daniel Jacob

INRA UMR 1332 BFP – Equipe "Métabolisme"

Plateforme Métabolome Bordeaux

INRA SCIENCE & IMPACT

Capture des données … **le plus en amont possible**

Permet de capturer toutes les données

Evite ainsi de faire de l'archéologie de données
au moment de leur diffusion, de leur publication

Données publiées en étroite relation avec le message
=> **Malheureusement, seule une partie des données est concernée**

Capture des données … **le plus en amont possible**

Permet de capturer toutes les données

Evite ainsi de faire de l'archéologie de données
au moment de leur diffusion, de leur publication

Données publiées en étroite relation avec le message
=> **Malheureusement, seule une partie des données est concernée**

Comment encourager les producteurs de données à **investir du temps et de l'énergie**
en décrivant puis en formatant toutes leurs données
en amont de l'étape d'analyse des données?

Capture des données … **le plus en amont possible**

Permet de capturer toutes les données

Evite ainsi de faire de l'archéologie de données
au moment de leur diffusion, de leur publication

Données publiées en étroite relation avec le message
=> **Malheureusement, seule une partie des données est concernée**

Comment encourager les producteurs de données à **investir du temps et de l'énergie**
en décrivant puis en formatant toutes leurs données
en amont de l'étape d'analyse des données?

Inciter plutôt que contraindre

Proposer des services qui leur soient vraiment utiles,
en leur permettant de gagner en efficacité là où ils aimeraient en gagner

Tenir compte de leur mode de fonctionnement, de leurs habitudes de travail

Use-Case "Metabolism"

Plant Metabolism

Systems Biology - Biomarkers associated with plant performance

Several partners

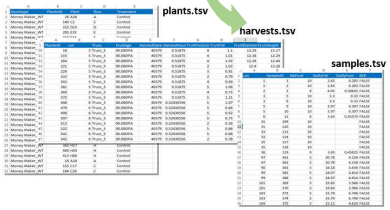
Sample identifiers

Several operators, technics, data types, SOPs, …

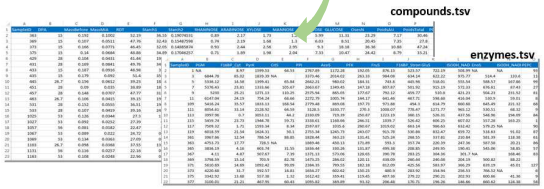
Experiment Design

Raw data

Experiment Data Tables

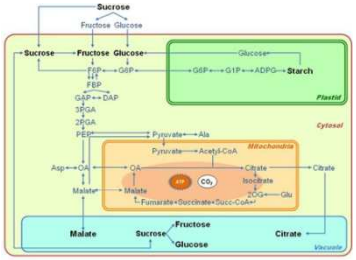


Data + Metadata



Annotation, Curation, Validation

Modelisation

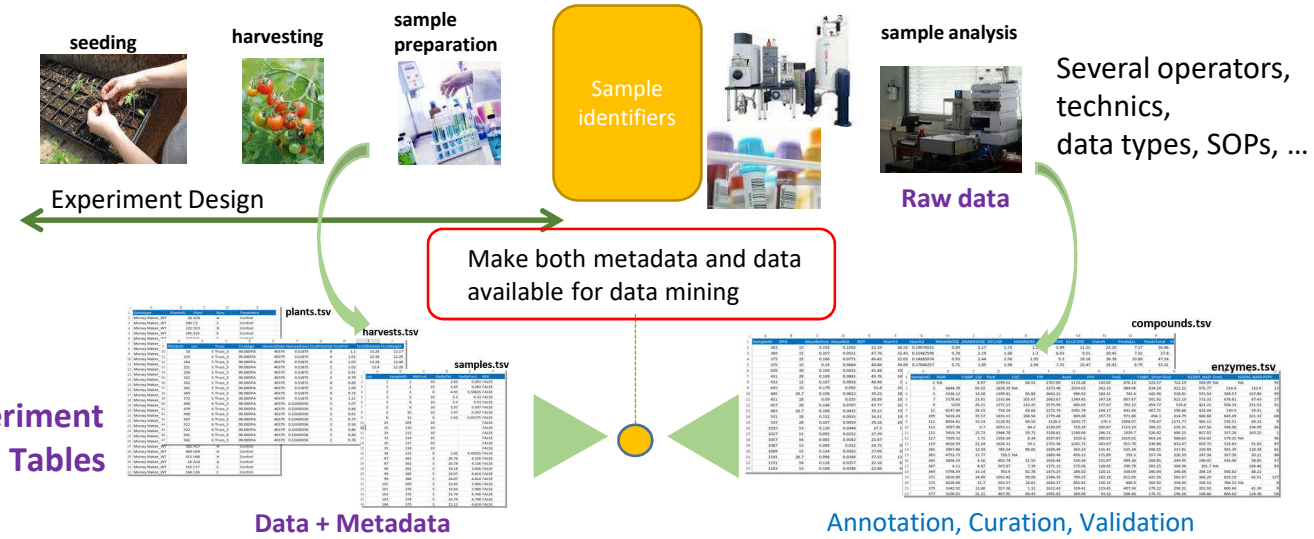


Use-Case "Metabolism"

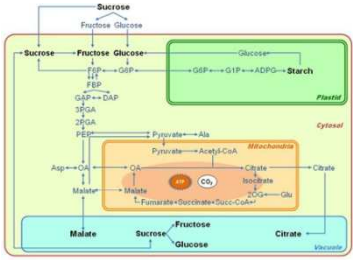
Plant Metabolism

Systems Biology - Biomarkers associated with plant performance

Several partners



Modelisation



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**

Classically, these two types of tasks must be supported

Capture des données … le plus en amont possible

Tenir compte de leur mode de fonctionnement, de leurs habitudes de travail

Utilisation du tableur comme outil central

Malgré tous leurs inconvénients

par ex: information multiple dans un format sans structure interne

Cela n'enlève rien à leurs avantages

outil universel

Capture des données … le plus en amont possible

Tenir compte de leur mode de fonctionnement, de leurs habitudes de travail

Utilisation du tableur comme outil central

Malgré tous leurs inconvénients
par ex: information multiple dans un format sans structure interne
Cela n'enlève rien à leurs avantages
outil universel

⇨ **Tâches répétitives et fastidieuses**

Collecte et préparation des données :

- beaucoup de manipulations de données, essentiellement sous forme de tableaux,
- combiner des ensembles de données en fonction d'un champ commun (identifiants)

Modélisation :

- sélection de sous-ensemble de données puis nombreuses répétitions de traitements complexes selon un paramétrage très varié (scénarios).

Capture des données … le plus en amont possible

Tenir compte de leur mode de fonctionnement, de leurs habitudes de travail

Utilisation du tableur comme outil central

Malgré tous leurs inconvénients
par ex: information multiple dans un format sans structure interne
Cela n'enlève rien à leurs avantages
outil universel

⇨ **Tâches répétitives et fastidieuses**

Collecte et préparation des données :

- beaucoup de manipulations de données, essentiellement sous forme de tableaux,
- combiner des ensembles de données en fonction d'un champ commun (identifiants)

Modélisation :

- sélection de sous-ensemble de données puis nombreuses répétitions de traitements complexes selon un paramétrage très varié (scénarios).

Permettre de gagner en efficacité là où ils aimeraient en gagner

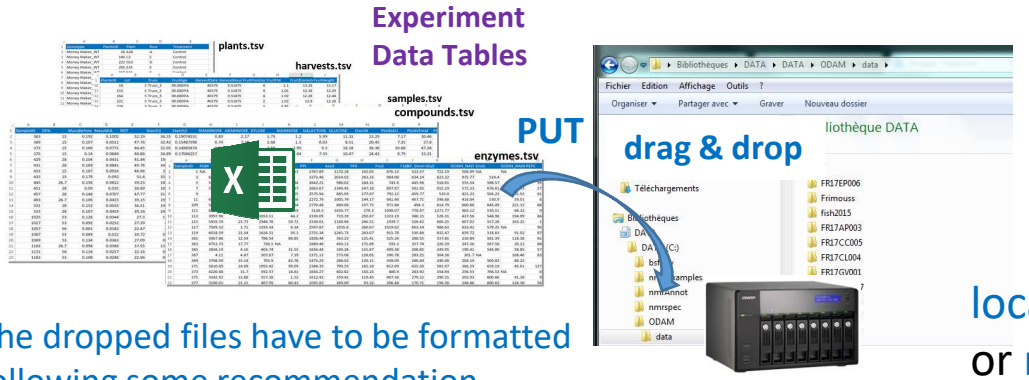
Prise en charge de l'ensemble des tâches relatif à **la gestion et la manipulation de données**:

- Fusion, sélection de sous-ensemble,
- Exploration (visualisation) multicritère,
- Stockage puis partage avec leurs partenaires …

ODAM Framework
Open Data for Access and Mining

Data capture

ODAM Framework
Open Data for Access and Mining



The dropped files have to be formatted following some recommendation

local (*i.e. User Disk, or a NAS*)
or remote (*virtual disk space*).

The central idea

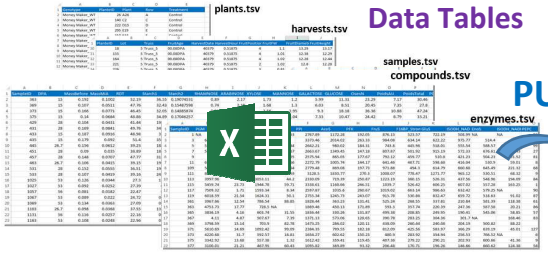
data producers have to "*just drag & drop*" their data tables on a storage space

... to gain access to services

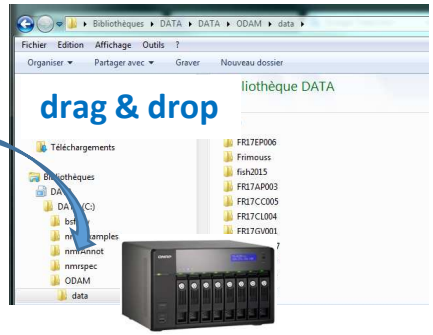
Data capture

ODAM Framework
Open Data for Access and Mining

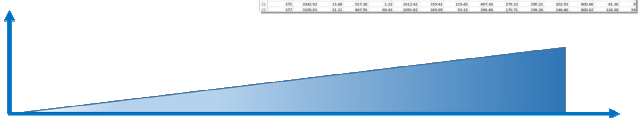
Experiment
Data Tables



PUT



Services offered to Data Producers/ Consumers



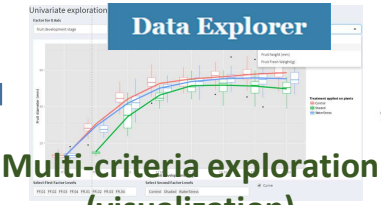
data can be dropped as soon as they are available

Storage and sharing with all partners

GET

Using Data

Merging & selection of data subsets



Multi-criteria exploration (visualization)

Data can be downloaded, explored and mined

Repetition of complex treatments according to very varied parameters

R package ***Rodam***

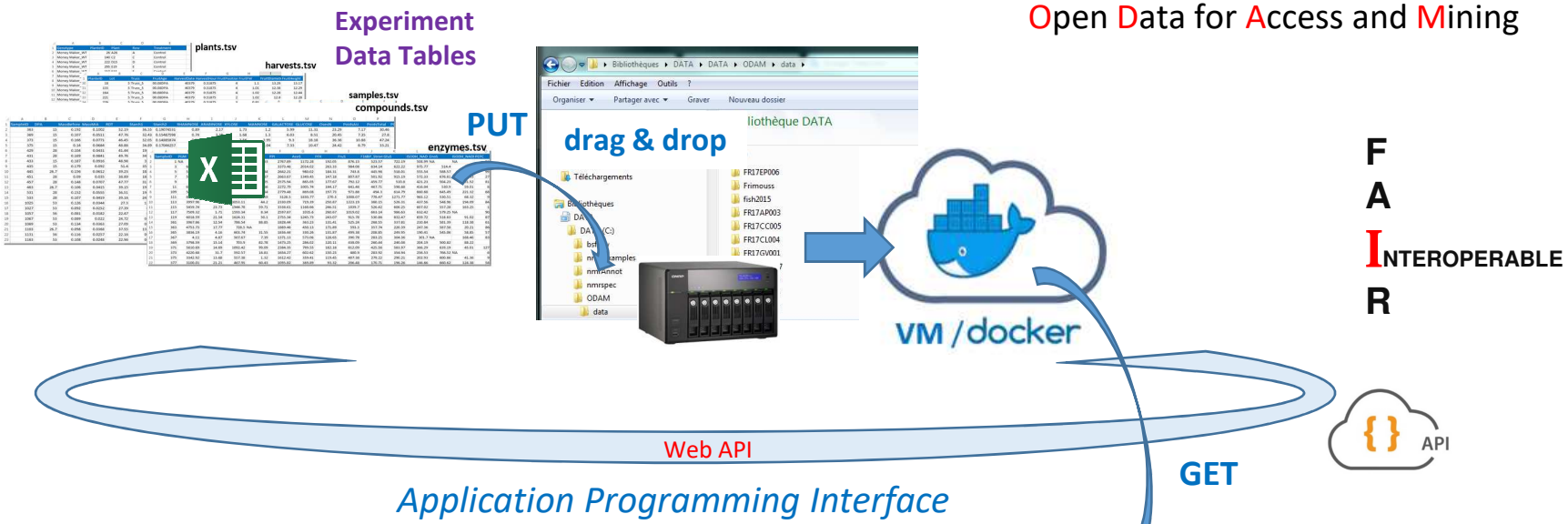
CRAN 0.1.4

Develop if needed, lightweight tools

- R scripts (Galaxy), lightweight GUI (R shiny)

Data capture

ODAM Framework
Open Data for Access and Mining



Using Data

All these services are **based on the API layer**, which **ensures interoperability** between the different tables and the applications that enable them to be used.

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

Data can be downloaded, explored and mined

Documented API along with online tool for developers

swagger

Proposer une assistance

Des **outils / services** même très pertinents **ne suffisent pas** à répondre par eux-mêmes à la capture des données.

- Il faut mettre en place des guides de bonnes pratiques simples, compréhensibles et auxquelles les producteurs de données peuvent facilement adhérer.
- Proposer une assistance afin de formater leurs données.
- Il faut les inciter à faire la partie du chemin qui nécessite leurs connaissances/expertises du domaine et de leurs données,
- Faire à leur place l'autre partie du chemin (formatage finalisé)

Proposer une assistance

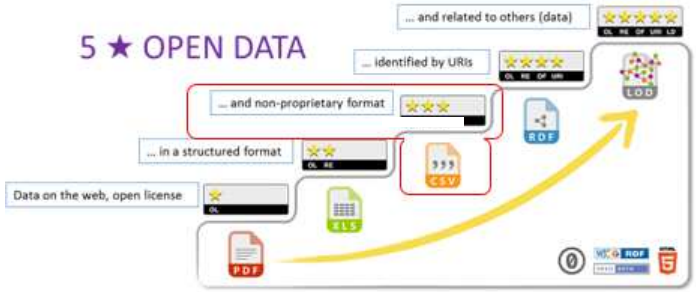
Des **outils / services** même très pertinents **ne suffisent pas** à répondre par eux-mêmes à la capture des données.

- Il faut mettre en place des guides de bonnes pratiques simples, compréhensibles et auxquelles les producteurs de données peuvent facilement adhérer.
- Proposer une assistance afin de formater leurs données.
- Il faut les inciter à faire la partie du chemin qui nécessite leurs connaissances/expertises du domaine et de leurs données,
- Faire à leur place l'autre partie du chemin (formatage finalisé)

Utilisation du tableur comme outil central

Néanmoins

Promouvoir des formats non propriétaires comme CSV ou TSV



étape nécessaire et indispensable vers le « **Linked Open Data** ».

Promote good practices

Experiment Design (DoE)

Use-Case "Metabolism"

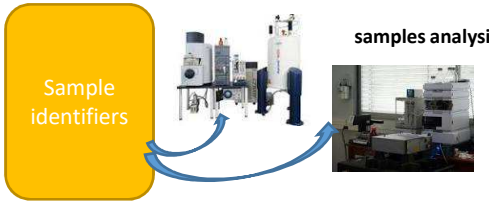
samples : Sample features

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

Promote non-proprietary format like CSV or TSV

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					
512	Control	FF.03	21DPA	NA					
117	Control	FF.01	07DPA	3					
536	Control	FR.02	47DPA	NA					
544	Control	FR.03	50DPA	NA					
158	Control	FF.04	35DPA	5					
109	Control	FF.03	22DPA	7					
134	Control	FF.02	15DPA	3					
31	Control	FF.01	08DPA	4					
179	Control	FF.03	28DPA	3					
383	Control	FF.04	34DPA	5					
425	Control	FR.04	55DPA	2					
520	Control	FF.03	30DPA	NA					
419	Control	FR.03	50DPA	2					
138	Control	FF.02	15DPA	6	27.96	22.14	9.69	A	T7
143	Control	FF.03	29DPA	4	48.45	42.92	51.35	A	T6
365	Control	FR.02	47DPA	5	55.11	44.9	71.82	A	T6
127	Control	FF.03	27DPA	3	45.71	43.28	47.8	A	T5
188	Control	FR.01	42DPA	3	55.38	47.1	77.39	A	T6

Use-Case "Metabolism"

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

Metadata

Promote non-proprietary format like CSV or TSV

Data

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

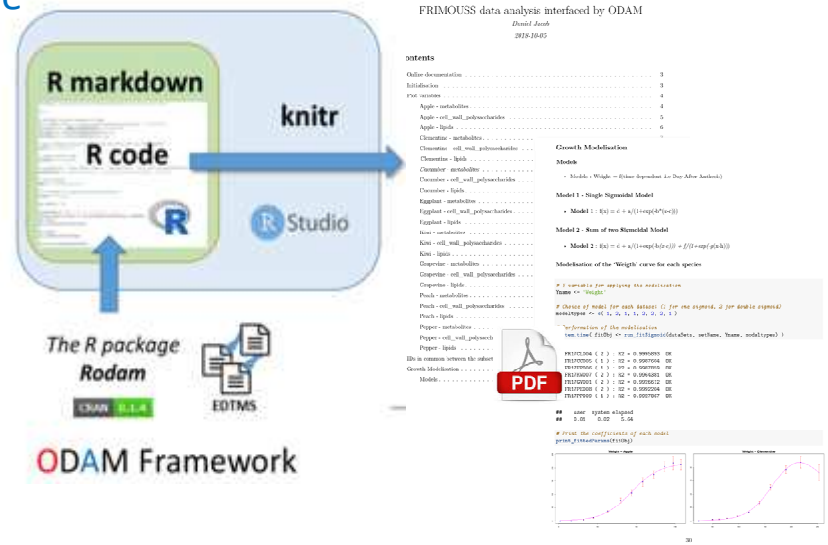
Opportunité vers la Recherche Reproductible

F
A
I
R E-USABLE

Reproduire les résultats présentés dans un article.

Favorisée par le fait que l'ensemble des données avant analyse est disponible via des langages de scripts (R, API)

Collecte des données et préparation des données pour l'analyse effectuées par l'approche ODAM



R package ***Rodam*** CRAN 0.1.4

Vignettes: Wrapper Functions for ODAM (Open Data for Access and Mining) Web Services

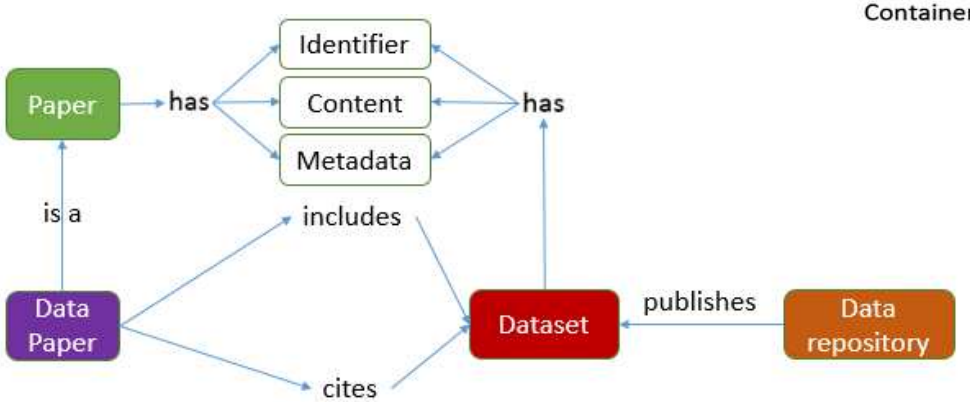
rescience.github.io

Opportunité vers l'Open Data

Diffusion des données

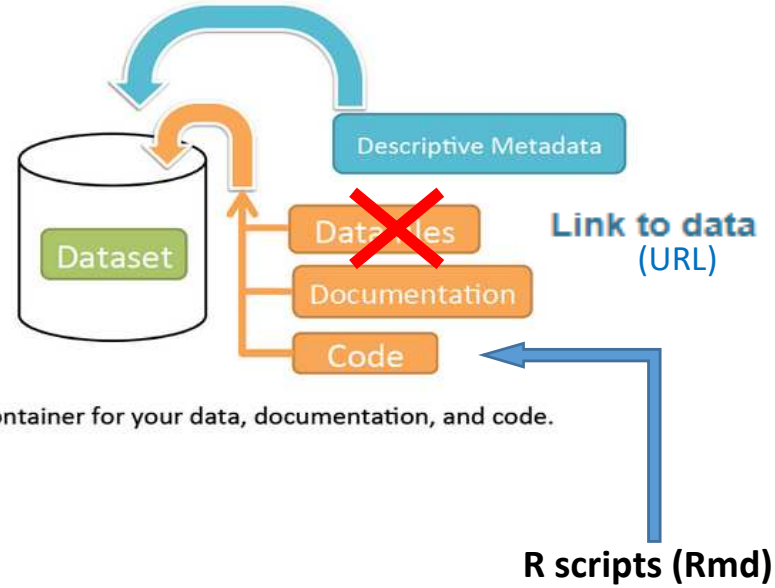
Diffuser plus rapidement leurs données
au moment qu'ils le souhaitent
sans aucun effort supplémentaire

Data paper



F INDABLE
A CCESSIBLE
I
R

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



ODAM Framework
Open Data for Access and Mining

https://fr.slideshare.net

"Make your data great now" vers l'Open Data et la Recherche Reproductible

"Make your data great again" vers le Linked Open Data

ODAM Framework
Open Data for Access and Mining

https://fr.slideshare.net

"Make your data great now" vers l'Open Data et la Recherche Reproductible

"Make your data great again" vers le Linked Open Data

Merci de votre attention