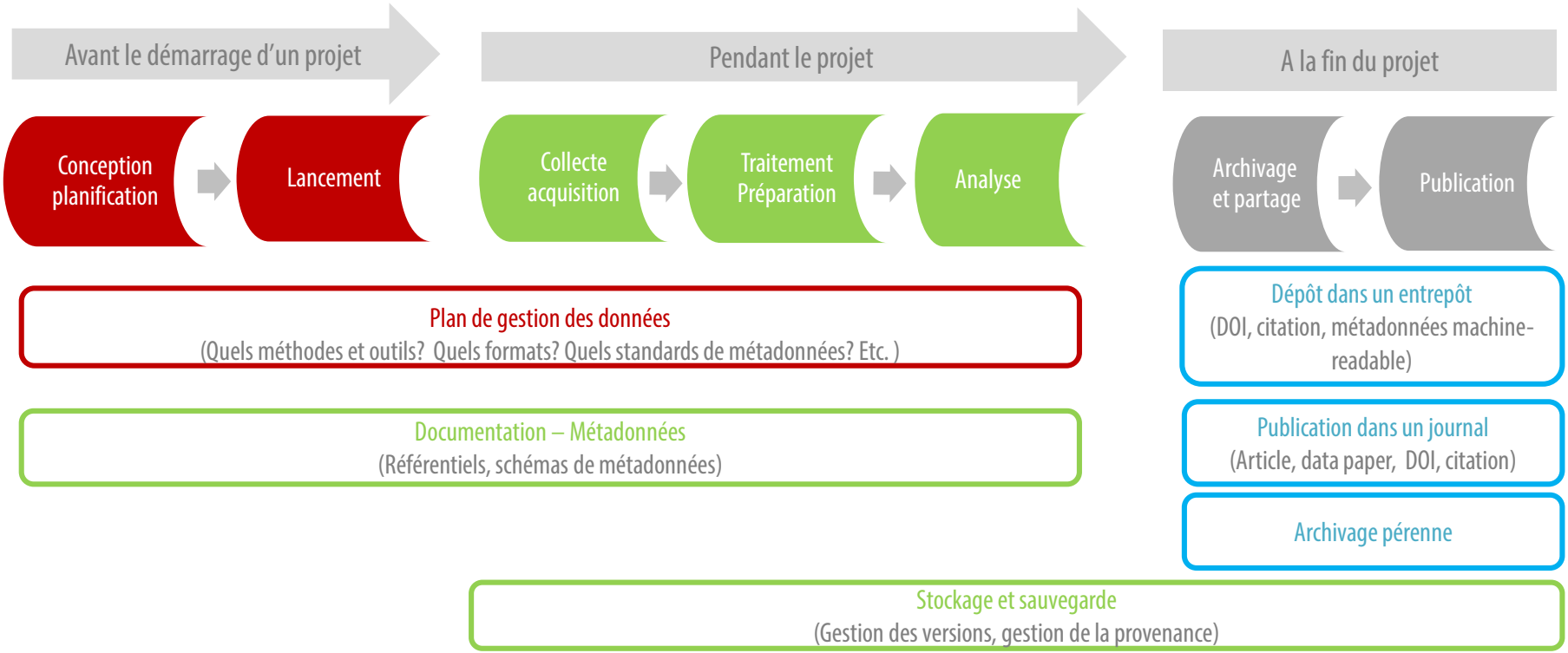
Comment mettre à disposition ses données?

data2019opensci : En route vers la science ouverte : le cycle de vie des données

NOM DE Esther Dzalé Yeumo

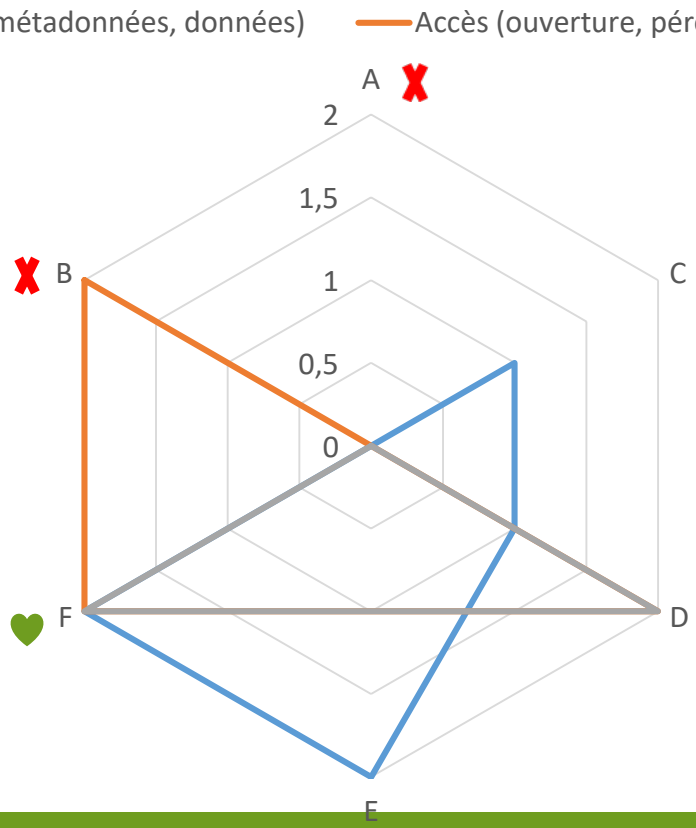
FAIR BY DESIGN



Réfléchir à une stratégie globale

Voies de diffusion des données	Avantages 	Limites & freins 
Données déposées dans un entrepôt	• Nombreux entrepôts connus et reconnus de la communauté scientifique • Indexation par des moteurs de recherche (Google data search) • Citation directe des données possible	Selon le choix de l'entrepôt o Métadonnées plus ou moins riches (réutilisation plus ou moins aisée) o Diffusion parfois orientée vers une communauté spécifique, ce qui peut limiter la visibilité des données hors la communauté cible
Données intégrées dans un article classique	• Intégration maximale des données et de l'article : citable, recherchable	• Données difficiles à trouver indépendamment de l'article et dans une forme peu ou pas réutilisable • Absence d'information sur les données brutes et les traitements ayant abouti aux données intégrées dans l'article → limite la reproductibilité
Données en matériel supplémentaire d'un article classique	• Format des données libéré des contraintes de rédaction de l'article (format, volume, nature des données…)	• Taille (poids de fichier) souvent limitée • Données plus difficiles à trouver indépendamment de l'article et dans une forme peu ou pas réutilisable, présentation hétérogène
Données déposées dans un entrepôt et associées à un article (data paper ou article classique)	• **Visibilité accrue** : via entrepôt et article, citations croisées • **Recherche et réutilisation des données facilitée** par la richesse des métadonnées, en particulier si data paper • **Peer Review, crédit aux auteurs**	• *Choix de l'entrepôt pour le jeu de données : disciplinaire (biologie, sciences de la vie, sciences du sol, chimie), institutionnel (Data Inra), générique (Zenodo …)* • Temps de rédaction supplémentaire pour data papers • Coût de la publication

Choisir le bon dispositif



A. Données et métadonnées non structurées + accès fermé et non pérenne + absence visibilité
B. Accès ouvert et pérenne mais absence structuration données et métadonnées et zéro visibilité
C. Métadonnées ou données structurées + Accès fermé et non pérenne + zéro visibilité
D. Métadonnées structurées + accès ouvert et pérenne + visibilité accrue
E. Métadonnées et données structurées + accès ouvert mais non pérenne + visibilité réduite
F. Métadonnées et données structurées + accès ouvert et pérenne + visibilité accrue

La citation indispensable pour induire un cercle vertueux

Figure 1: Data citation example.

...highly site specific, potentially limiting their wider value. However, applying the approach as conducted in this paper to data such as that presented by Barnett et al (2013) to (1) ve relative values for different organisms should provide a more generic set of reference data'. In taking the REML approach forward it will be beneficial to target...

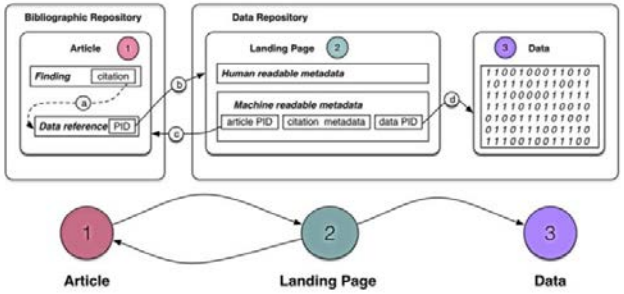
References

Barnett et al., 2013) C (2) arnett, N.A. Beresford, L.A. Walker, M. Baxter, C. Wells, D. Copplestone
Element and radionuclide concentrations in representative species of the ICRP's reference animals and plants and associated soils from a forest in North-west England.
NERC - Environmental Information Data Centre (3) http://doi.org/10.5285/e40b53d4-6699-4557-bd55-10d196ece9ea

(1) Data citation in text; (2) Reference; (3) Globally resolvable unique identifier. *Example from Beresford NA, et al. (2016). Available at* https://doi.org/10.1016/j.jenvrad.2015.03.022[16].

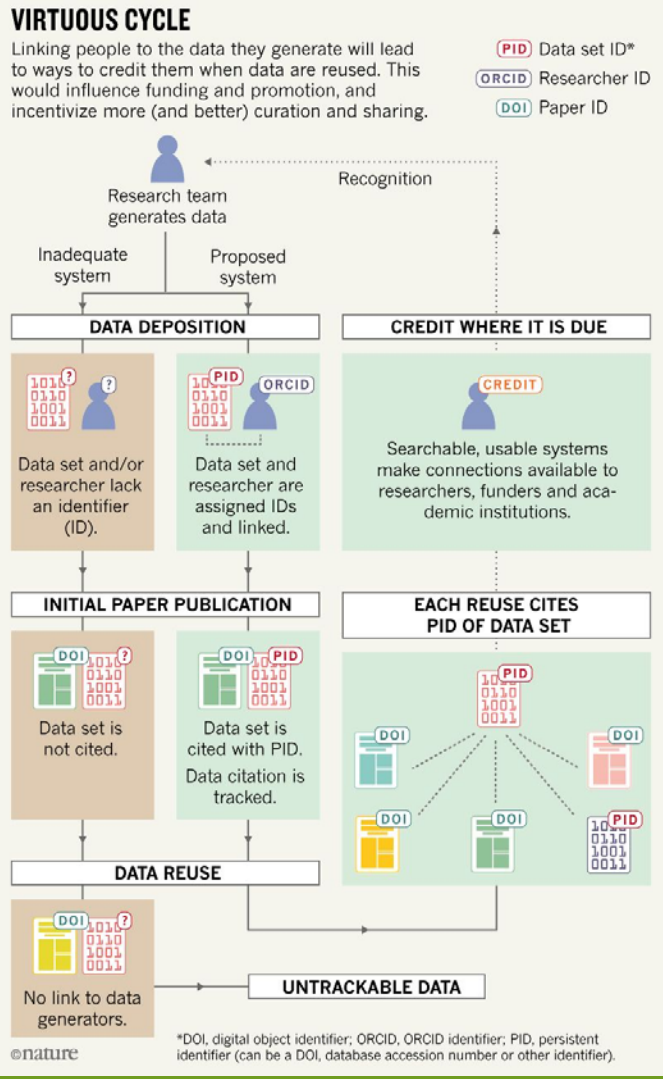
Full size image »

Figure 2: Data citation resolution structure (ideal workflow).



Articles (1) link to datasets in appropriate repositories, on which their conclusions are based, through citation to a dataset (**a**), whose unique persistent identifier (PID) resolves (**b**) to a landing page (2) in a well-supported data repository. The data landing page contains human- and machine-readable metadata, to support search and to resolve (**c**) back to the citing article, and (**d**) a link to the data itself (3).

Full size image »



https://www.nature.com/articles/sdata2018259 et
https://www.nature.com/articles/d41586-019-01715-4

INRA SCIENCE & IMPACT

Exemples concrets

Mise à disposition de données dynamiques

Situation avant

- Données de 4 grands lacs depuis plus de 50 ans accessible via un SI en ligne + de nombreux partenaires qui veulent tous être cités
- 2 types de demande :
  - Accès au SI
  - Demandes d'extraction de données spécifiques. Ex : 50 ans de données de chlorure sur le lac Léman

Pratique actuelle

- DOI sur le SI en tant qu'objet scientifique
- Dépôt des extractions dans Data Inra + DOI + citation
  - Plus d'opportunité d'être cité directement

Dépôt dans un entrepôt généraliste compatible FAIR

Estimation réalisée avec le **FAIR self-assessment tool** de l'ARDC

- De facto : attribution DOI ; Moteur de recherche + indexation Google data search, re3data, Fairsharing
- Effort utilisateur : fournir suffisamment de métadonnées

- De facto : Accès API, Protocole HTTP, Métadonnées toujours disponibles; métadonnées et certaines données structurées
- Effort utilisateur : préciser les conditions d'accès aux données ; choix des formats

- De facto : Métadonnées conformes standards, machine readable; conversion formats propriétaires vers formats ouverts
- Effort utilisateur : structuration des fichiers et choix des formats; vocabulaires standards; documentation suffisante

- De facto : Licence par défaut paramétrable; gestion de version et traçabilité des modif dans l'entrepôt; compatible PROV-O
- Utilisateur : fournir des métadonnées de provenance; Standards communautaires, …

Adaptation diapo Dimitri Dzabo

La proposition de RDA pour la mise à disposition de données dynamiques

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

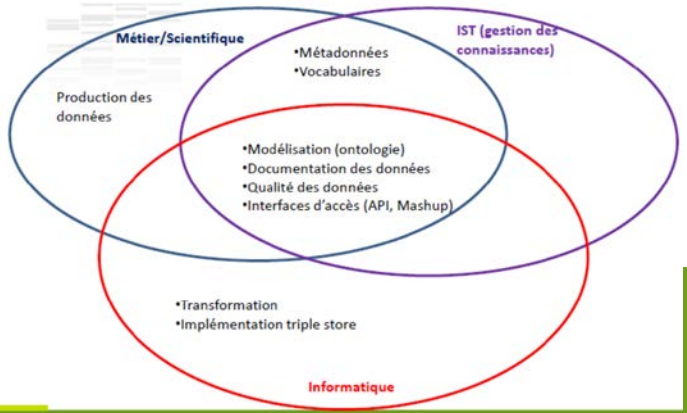
Le projet scientifique – ALTERPRO

INRA*, Plante&Cité, ONEMA, Plan Ecophyto 2018

- Contrôler les populations de Processionnaire du Pin
  - à un niveau tolérable permettant de protéger hommes, plantes et animaux
  - avec des moyens écologiques : piéger les papillons mâles avec des pièges à phéromones pour limiter la reproduction
- Comparer l'efficacité des pièges et des phéromones disponibles sur le marché

* UE Entomologie et Forêt Méditerranéenne

Compétences mobilisées



Anne-Sophie Brinquin
Jean-Claude Martin
Sophie Aubin
Sylvie Cocaud
Pascal Aventurier
Esther Dzalé

Nos objectifs, nos contraintes

- Objectifs
  - Démontrer par l'exemple l'intérêt des approches sémantiques
  - Tester les outils de structuration et de mise à disposition de données RDF
  - Estimer les efforts et les compétences nécessaires dans de telles approches
- Nos contraintes
  - Solliciter le moins possible les scientifiques
  - Démontrer une plus-value des approches sémantiques

Description des sites

Données géographiques

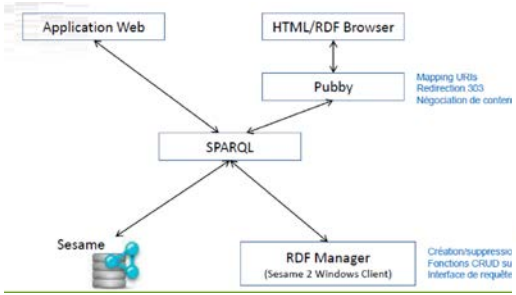
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Communes	Saint-Fargeau-Ponthierry	Obernai	Angers							Saint-Martin-de-Ré	Marenne
2	Description Sites Témoins	Type de configuraton	alignement	pas de site témoin déterminé	îlot boisé							îlot boisé	îlot boisé
3		Description des sites	parc de loisir		Echangeur de la Baumette							Parcelle forestière	à proximité la piscine Marenne
4		Géolocalisation	48°31'56.85"N, 2°33'47.42"E		47°27'53.61"N, 0°34'14.82"O							46°15'3.24"N, 1°31'0.58"O	45°49'28.2 1° 8'19.54
5		Surface	56 m		25100 m²							0,13 ha	10 ha
6		Historque du site	non communiqué		non communiqué							Comptage des nids par l'ONF depuis 2003 (données transmises)	non communiq
7		Nombre de pins et/ou de cèdres	20		62							116	non communiq
8	Sites Tests	Nombre de sites tests	1	1	7							1	1
9		Type de configuraton	alignement	îlot boisé	alignement	îlot boisé	alignement	alignement	alignement	îlot boisé	alignement	îlot boisé	îlot boisé
10		Description des sites	parc de loisir	jardin privé	Roseraie I bord de route	Roseraie II Parc	Roseraie III butte antibrut	Roseraie IV quartier résidentiel	Arboretum V Falun d'Orgemont	Arboretum VI cours d'école	Arboretum VII Plaine de jeux	parc	bosquet
11		Géolocalisation	48°31'59.30"N, 2°33'29.47"E	48°27'34.36"N, 7°30'18.04"E	47°26'55.62"N, 0°34'21.40"O	47°26'48.09"N, 0°33'51.29"O	47°26'33.64"N, 0°33'28.63"O	47°26'38.59"N, 0°33'14.78"O	47°26'43.17"N, 0°32'29.08"O	47°27'10.16"N, 0°32'47.32"O	47°27'45.99"N, 0°31'39.55"O	46°10'54.61"N, 1°22'26.48"O	45°49'14.1 1° 8'10.14
12		Surface	500 m	< 15 ares	3000 m²	17939 m²	8650 m²	8650 m²	3250 m²	5270 m²	1077 m²	4,6 ha	1 ha

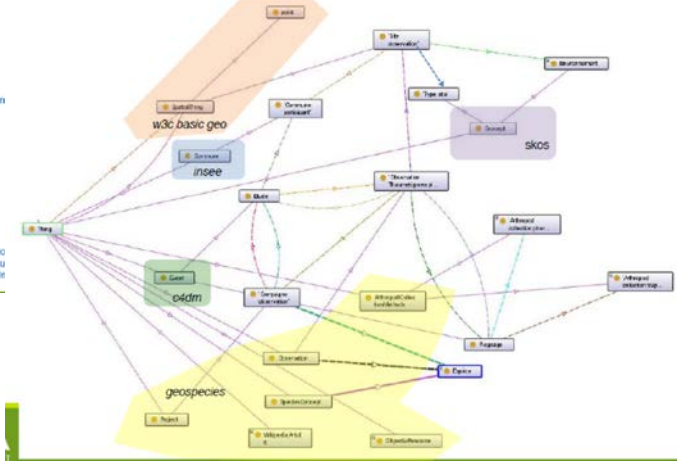
Protocole

Relevés

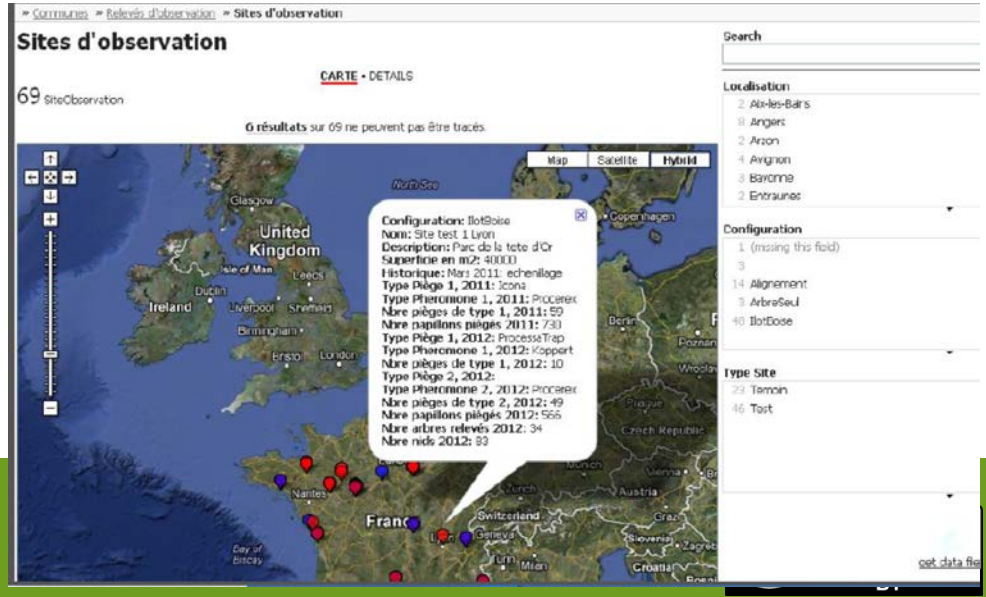
Dispositifs pièges à phéromones été 2011											
	Relevés des nids réalisés à l'hiver 2011	non réalisés	non réalisés	non réalisés							nor réalisés témc
	Nombre de pièges installés	10	15	14 a) 1 b)	9	12 a) 1 b)	12 a) 2 b)	8 a) 1 b)	6 a) 1 b)	8	18
	Type de pièges	Procerex	Mastrap	a) Mastrap b) Nufarm	Mastrap	a) Mastrap b) Nufarm	a) Mastrap b) Nufarm	a) Mastrap b) Nufarm	a) Mastrap b) Nufarm	Mastrap	Mast
	Type de phéromones	Procerex	Procerex	Procerex	Procerex	Procerex	Procerex	Procerex	Procerex	Procerex	Proce
	Date de la pose des pièges	25 juillet (sem 30)	01 juin (sem 22)	01 juillet (sem 26)			01 juillet (sem 26) et 06 juillet (sem 27)	31 mai (sem 22) et 19 juillet (sem 29)	19 juillet (sem 29)	19 juillet (sem 29)	12 ju (sem
	Premiers vols de papillons constatés	non communiqué	Sem 26 et 27	sem 25							non re
	Relevés intermédiaires réalisés	non réalisés	19 juin 4 juillet	non réalisés							30 ju (sem
	Nombre de papillons piégés au total	non communiqué	8	340	305	430	201	390	285	155	58

Bilan





- Les plus
  - Des données harmonisées, standardisées, désambiguïsées, contextualisées et enrichies (lien avec TaxonConcept, Geospecies, Event, FOAF, qudt, et Insee)
  - Des données publiées dans un triple store + une application Web pour visualiser les données enrichies
- Les moins
  - Une valorisation insuffisante des résultats
  - Seules des personnes averties savent où se trouvent ces données et peuvent aller les chercher
  - (Indépendamment de ce projet), pas vraiment d'outils d'analyse main stream intégrant des connecteurs aux entrepôts RDF



Le bon équilibre est le votre

- Comparer les coûts aux bénéfices escomptés
- « FAIR » au mieux, en fonction de son cas d'usage
  - Il peut être plus judicieux de mettre à disposition un CSV suffisamment documenté et identifié avec un DOI dans une infra favorisant sa découverte, son accès et sa préservation que de vouloir à tout prix publier du RDF dans une infra incertaine
- Choisir les outils adéquats
- Ne pas oublier la citation

Ne construisons pas des cimetières de données

- Des données FAIR
- Intégrer les entrepôts de données dans des e-infra, rapprocher données et calcul
- Proposer des services
- Attirer des experts pour contribuer à la chaîne de valeur des données (hackathons, collaborations)

Here lays my data

En résumé

La mise à disposition des données demande de l'anticipation, du temps et de l'énergie

Il est important de baser sa démarche sur les principes FAIR et de choisir les bons outils pour augmenter le potentiel de réutilisation des données

La citation est indispensable pour induire un cercle vertueux qui inclut la reconnaissance des acteurs du partage des données

Pourquoi mettre à disposition → comment mettre à disposition → comment maximiser le ROI

Merci pour votre écoute