« Pour commencer, pourriez-vous définir ‘données de la recherche’ ? »

Une tentative de réponse

Joachim Schöpfel, Eric Kergosien, Hélène Prost

L’obscur objet…

« Data is fuel of economy » (Kroes)

« Tout est donnée » (Chignard)

Mais alors – c’est quoi, exactement ? La question revient avec insistance…

Difficiles à définir à cause de leur grande variété, qu’elles fussent physiques ou numériques (Borgman)

Est-ce satisfaisant, pour la compréhension, mais aussi pour l’usage, la curation, la préservation ?

Le projet *D4Humanities*

Projet structurant : plusieurs laboratoires, SCD Lille 3, MESHS, Ecole Doctorale SHS, ANRT ; suivi par la Direction Recherche

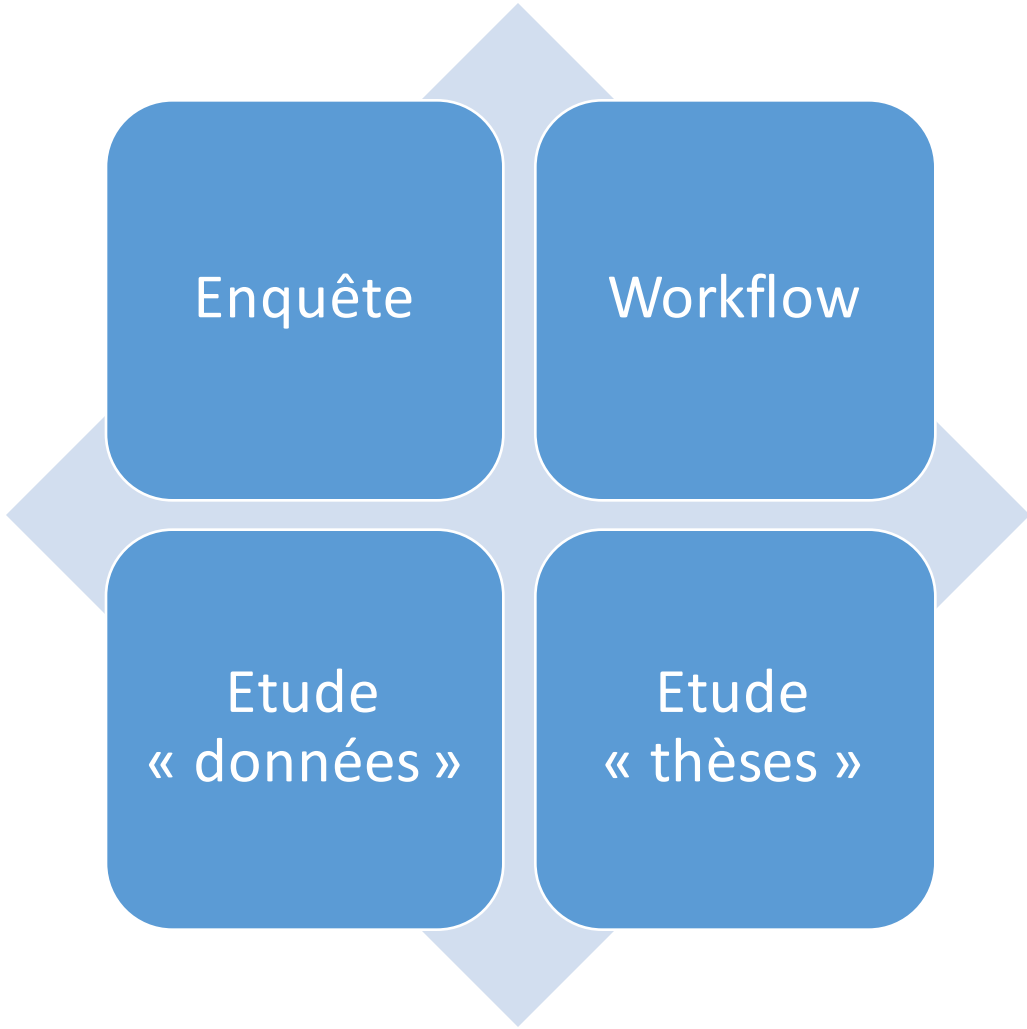
Projet émergent : préparation d'un projet franco-allemand ou européen (*xDiss*)

Modulaire : R&D

International : coopération avec l'ISN Oldenburg, NDLTD et ProQuest

Interdisciplinaire : SIC, linguistique

Transversal : partenariat recherche-métier



L'approche conceptuelle (1)

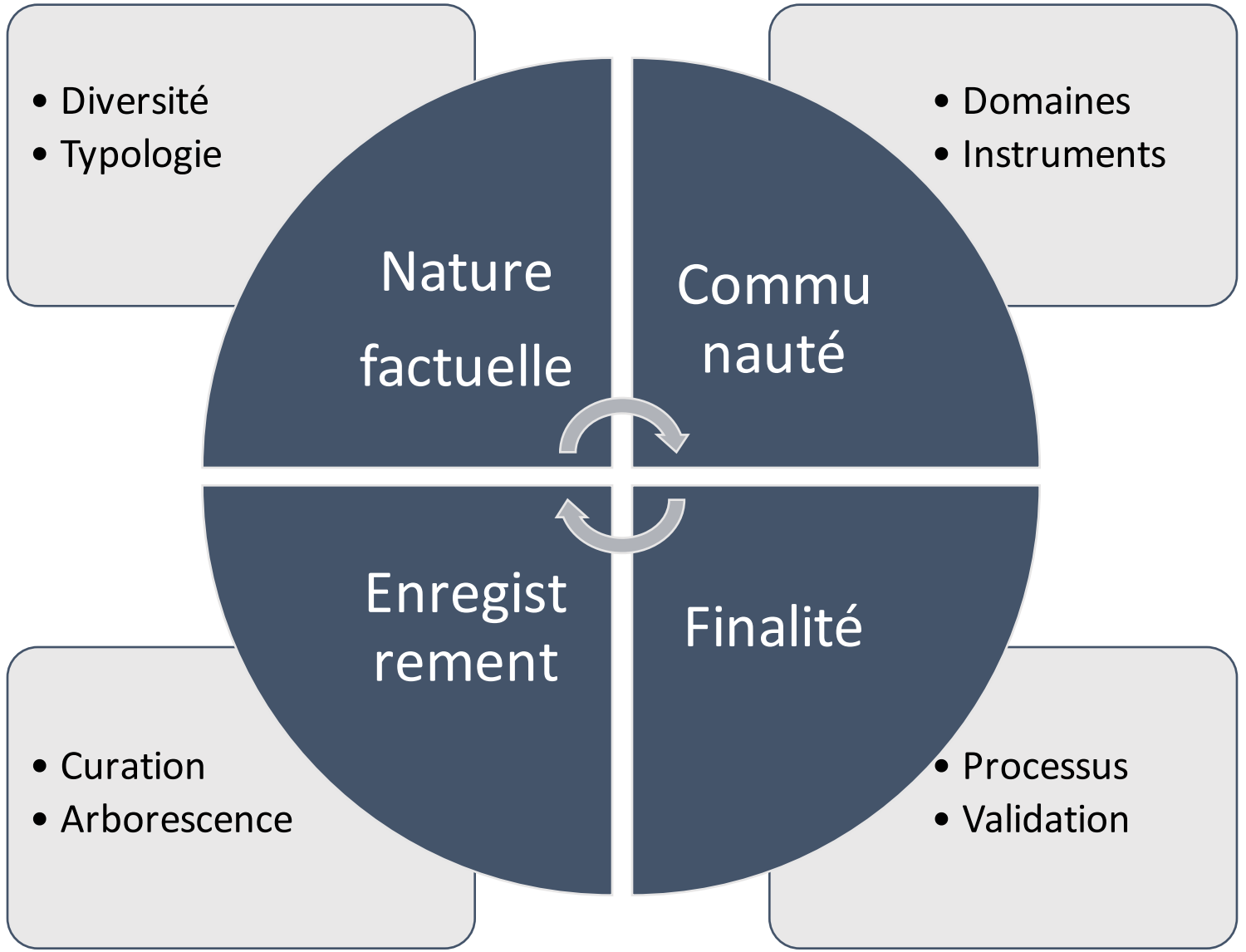
- Un concept aussi populaire que nébuleux et un état de l'art chaotique
- Beaucoup de définitions « implicites » faites d'anecdotes, de *success stories*, de descriptions, d'aspects technologiques, de tendances et d'impact sur les organisations et la société
- « Big Data is the Information asset characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value » (De Mauro et al. 2016)
- « What constitutes data is determined by a given community of interest that produces the data. However, an investigator may be part of multiple, overlapping communities of interest, each of which may have different notions of what are data » (Koltay 2016)

BIG DATA

Volume

Variété

Vélocité



L'approche conceptuelle (2)

Granularité ?

Collection ?

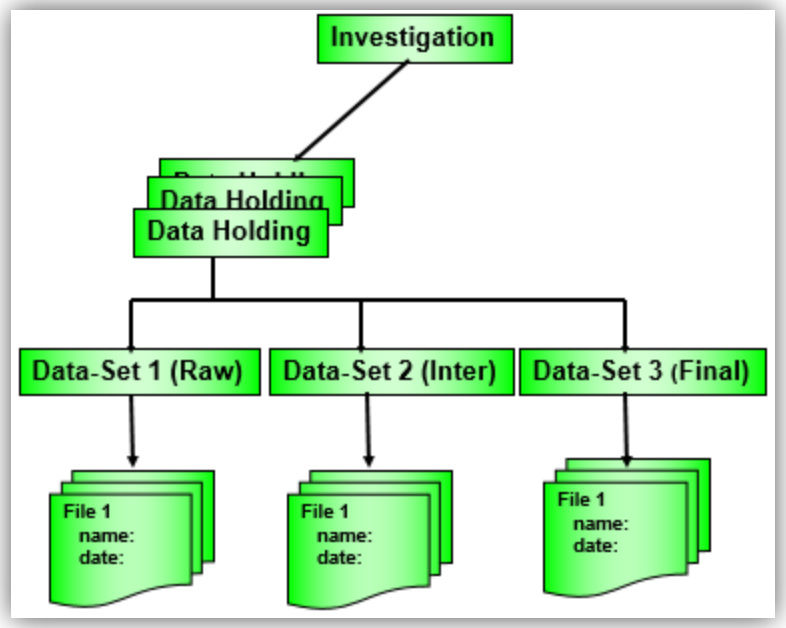
Ensemble ?

Identifiant ?

etc.

Souvent, l'aspect « gestion » est beaucoup plus important que les aspects « concept » ou « valeur »

- Pour le traitement (informatique, documentaire)
- Pour l'évaluation (plans de gestion, systèmes de recherche)



L’approche typologique (1)

« Data are most often defined by example, such as facts, numbers, letters and symbols » (Borgman et al. 2015)

Exemple d’une classification de données (RIN 2008 ; André 2015) :

- Les données d’observation ;
- les données d’expérimentation ;
- les données de simulation ;
- les données dérivées ;
- les données de référence.

Une synthèse en fonction des finalités et procédures de leur génération, davantage liée aux méthodes et outils de la recherche scientifique qu’aux disciplines et thématiques.

L'approche typologique (2)

Une répartition très inégale, avec plusieurs larges catégories, transversales aux disciplines, présentes dans une grande partie des domaines scientifiques, aux contours mal définis.

Une longue traîne d'autres types de données. D'après la catégorie *Other,* plus d'un tiers des sites indexés contient d'autres types de données, en dehors des quatorze catégories de la typologie.

Le rapprochement avec les disciplines montre que certains types de données sont surreprésentés, comme *Standard office documents*, *Plain text* et *Images* en SHS (Kindling et al., 2017).

	Count (n)	Percentage (%)
Scientific and statistical data formats	1152	63%
Standard office documents	1088	59%
Plain text	903	49%
Images	895	49%
Raw data	809	44%
Structured graphics	697	38%
Structured text	585	32%
Archived data	425	23%
Audiovisual data	339	18%
Software applications	324	18%
Databases	313	17%
Networkbased data	112	6%
Source code	81	4%
Configuration data	43	2%
Other	668	36%
Total	1837	100%

Types de données dans le répertoire re3data (N=1837 sites, 4 avril 2017)

L'approche typologique (3)

Nos études confirment une typologie propre aux SHS mais illustrent surtout l'intérêt d'une distinction entre données primaires (sources) et secondaires (résultats), avec une classification différente.

La correspondance entre ces catégories et les disciplines est forte, sans que l'on puisse parler de données spécifiques aux disciplines.

Les analyses attestent davantage de profils disciplinaires pour les différents types de données, voire de profils de données pour certaines disciplines

	re3data	Prost & Schöpfel 2015, sources	Prost & Schöpfel 2015, résultats
Scientific and statistical data formats	63%	26%	49%
Standard office documents	59%		
Plain text	49%	64%	76%
Images	49%	25%	21%
Raw data	44%		
Structured graphics	38%		32%
Structured text	32%		
Archived data	23%	34%	
Audiovisual data	18%	6%	44%
Software applications	18%		9%
Databases	17%		37%
Networkbased data	6%		
Source code	4%		
Configuration data	2%		
Enquêtes et entretiens		47%	
Observations		41%	
Expériences		36%	
Cartes et plans			10%
Other	36%	7%	3%
Total	100%	100%	100%

Classification des données primaires et secondaires en SHS (en %)

Approche fonctionnelle (1)

Politique

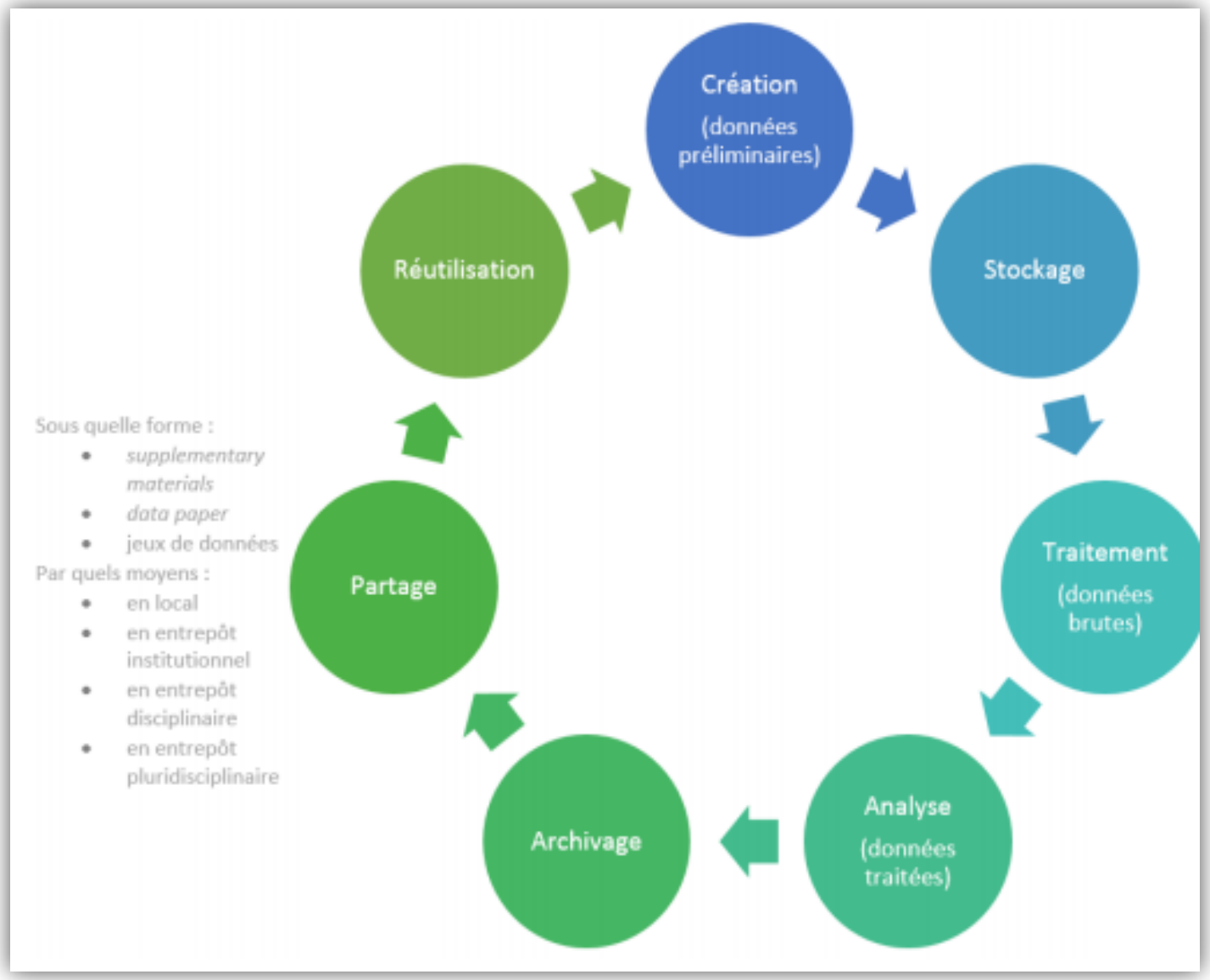
- Augmenter la transparence
- Créer un environnement favorable à l'économie
- Rendre l'action publique plus efficace

Economique

- Optimiser la recherche
- Accélérer l'innovation (santé, environnement)

Scientifique

- Explorer
- Visualiser
- Comparer et/ou vérifier des résultats
- Valider des hypothèses



Cycle de vie des données de la recherche (Pain 2016, p.18)

Approche fonctionnelle (2)

- Lien avec la « production scientifique »
- Lien aussi avec l'administration et le financement de la recherche (H2020)
- Il serait utile de différencier données scientifiques et données publiques, comme le précise le COMETS (2015).
  - Les données scientifiques produites sur des fonds publics ont dans la majorité des cas vocation à devenir publiques.
  - Les données publiques ont vocation à devenir scientifiques lorsqu'elles concernent l'environnement, le climat, la santé ou encore l'aménagement du territoire.
  - Un exemple: les données disponibles sur le portail de la métropole européenne de Lille (MEL) , notamment les données géolocalisées, sont utilisées par les chercheurs du projet LivreEtlecture pour l'analyse des pratiques des citoyens dans et à l'extérieur des bibliothèques.

Approche fonctionnelle (3)

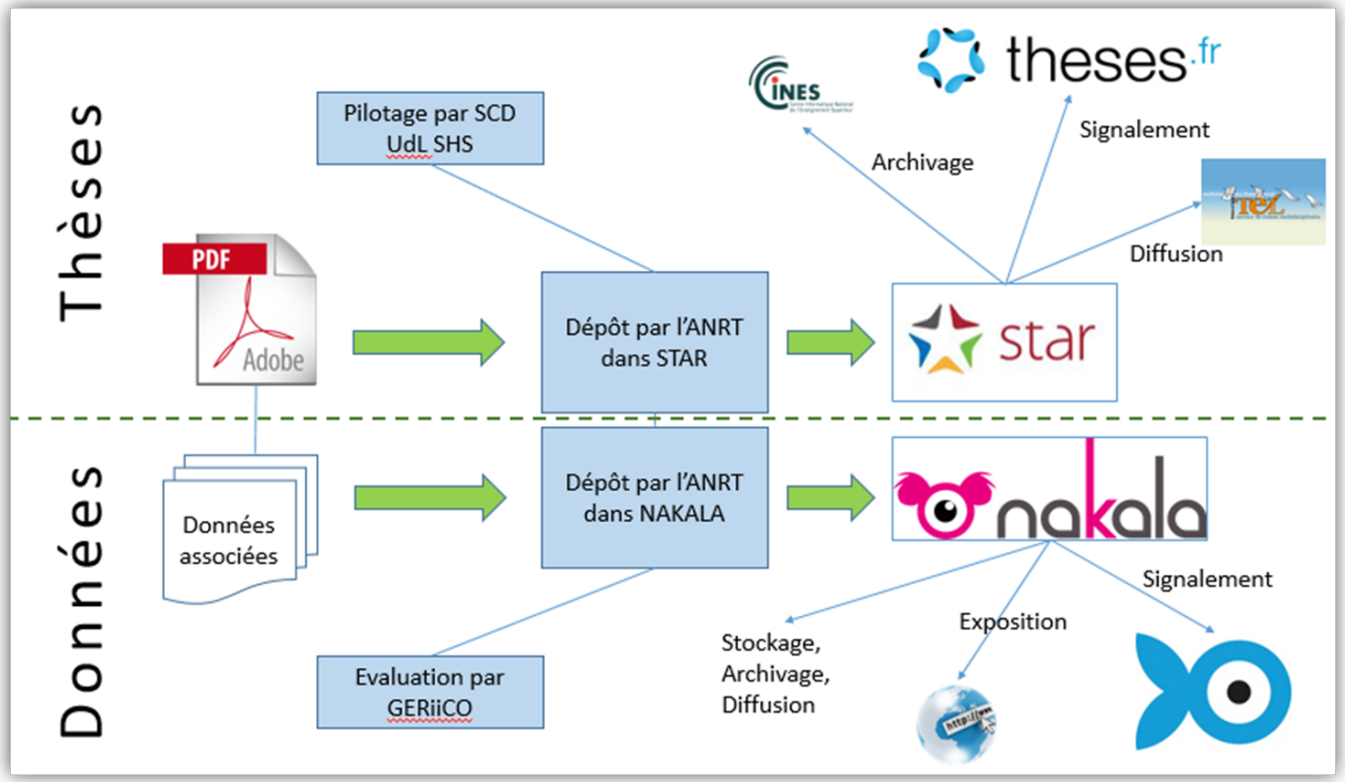
- Gestion, préservation et/ou partage ?
  - Lien avec le Big Data : « l'association dans une même analyse de données variées pour en déduire des informations que l'on n'était pas en mesure de trouver avec les analyses classiques de données structurées (…) souvent pour prendre une action en temps réel » (Cointot & Eychenne 2014, p.221)
- La gestion et mise à disposition des données scientifiques peuvent faciliter l'émergence de nouvelles formes d'analyses, avec des résultats novateurs, du simple fait de leur masse (volume) et leur diversité (variété), peut-être aussi (mais pas nécessairement) leur interopérabilité.

Objectifs D4Humanities

Conceptualiser plus finement le terme de données de la recherche dans le champ des SHS,

Par une définition des principaux types et niveaux de données et une problématisation de la distinction entre données primaires et secondaires.

Mais aussi, d'une manière opérationnelle, préparer la mise en place d'un workflow pour le dépôt des données des doctorants dans le TGIR Huma-Num.



Perspectives D4Humanities

Partenaires

- Université Humboldt de Berlin (projet eDissPlus)
- Karlsruhe Institut of Technology (projet bwDataDiss)
- Bielefeld University (projet CONQUAIRE)
- DANS aux Pays Bas

Analyse des travaux de re3data, CODATA, DataCite, DARIAH et CESSDA

Environnement *Open Science* (publication, réutilisation, évaluation)

Il s'agit en premier lieu d'une analyse de dispositifs et d'outils existants et d'exploiter les résultats d'autres recherches, y compris par une nouvelle analyse de nos propres résultats de 2015 (données dans les thèses, enquête sur le campus de l'Université de Lille SHS).

L'analyse s'appuiera notamment sur les 480 archives avec des données SHS répertoriées par re3data.

Livrables : colloque (thèses), rapport

Merci

D4Humanities est financé par la MESHS et par le Conseil Régional Hauts-de-France

Contact : joachim.schopfel@univ-lille3.fr