

Hors norme ?

Une approche normative des données de la recherche

Joachim Schöpfel

Introduction

Le cadre

- Code de la Recherche, Article L112-1
 - La recherche publique a pour objectifs : (...) e) L'organisation de l'accès libre aux données scientifiques.
- Plan d'action national 2018-2020, Engagement 18
 - Généraliser progressivement via un accompagnement la mise en place de plans de gestion des données dans les appels à projets de recherche, et inciter à une ouverture des données produites par les programmes financés
- Communiqué G7 septembre 2017
 - G7 Science Ministers committed to giving incentives for open science and to providing research infrastructures on the basis of FAIR data
- The First French GO FAIR Meeting For Future INs mars 2018
 - The rationale for prioritizing machine readable metadata over simple text

On parle de quoi ?

Simon Chignard :

« Tout est donnée »

« Un fait brut, qui n'est pas – encore – interprété »

Christine Borgman :

« Il est impossible de s'accorder sur une seule définition, en particulier en sciences humaines et sociales »

« *inputs, outputs, and assets of scholarship* »

Neelie Kroes :

« *Data is fuel of economy* »



Concept populaire mais nébuleux

- 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
- « (...) des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche » (OCDE 2007)
- « Qu'il s'agisse de leurs définitions, difficiles et multiples, de leur distinction pas toujours évidente d'avec les publications, de la difficulté à séparer parfois données collectées et données produites, de l'entremêlement des phases du cycle des données, de leur immense variété... : les données de recherche en SHS ne se laissent pas aisément définir et saisir » (Serres et al. 2017)

BIG DATA

Volume

Variété

Vélocité

- Diversité
- Typologie

- Domaines
- Outils

Nature
factuelle

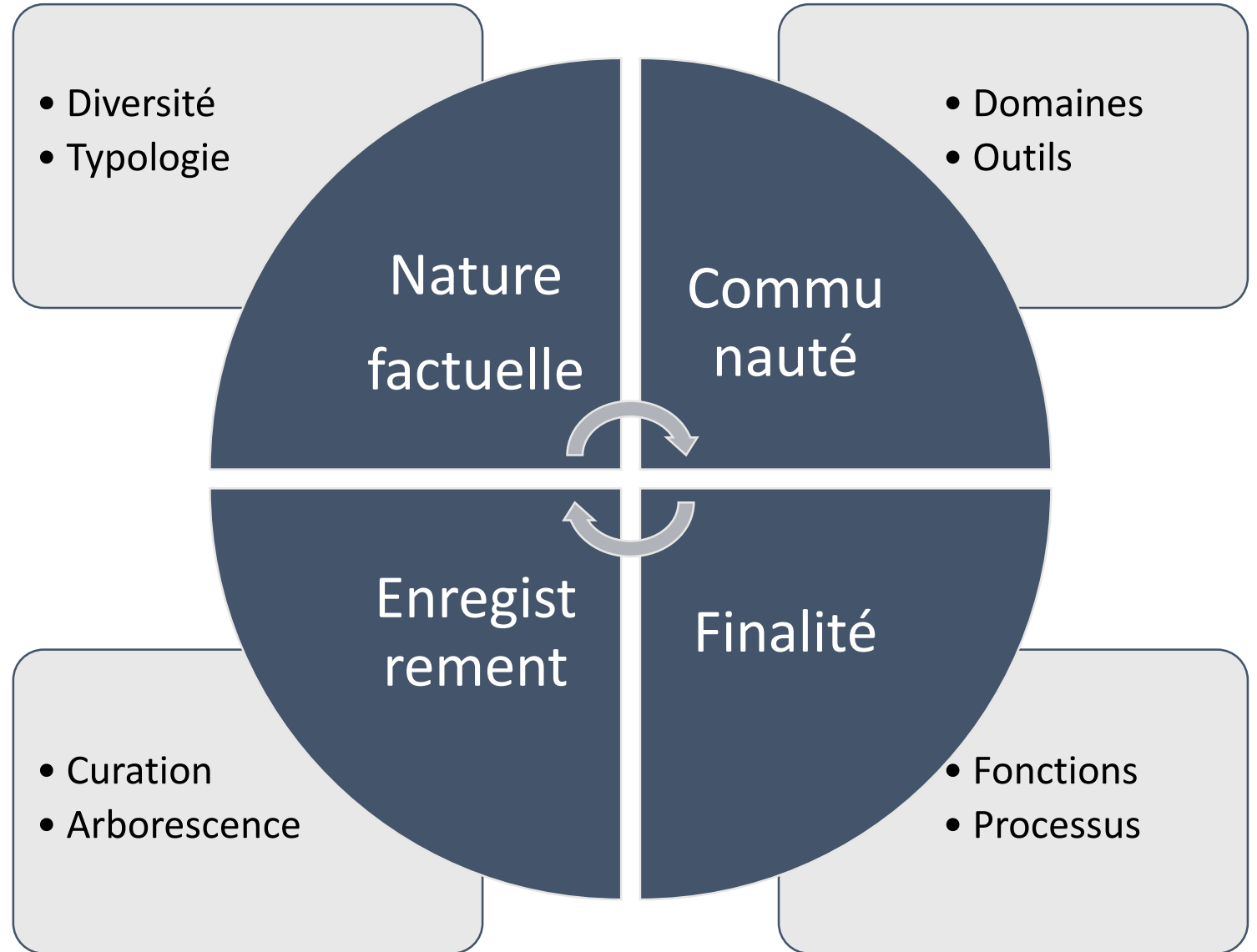
Commu
nauté

Enregist
rement

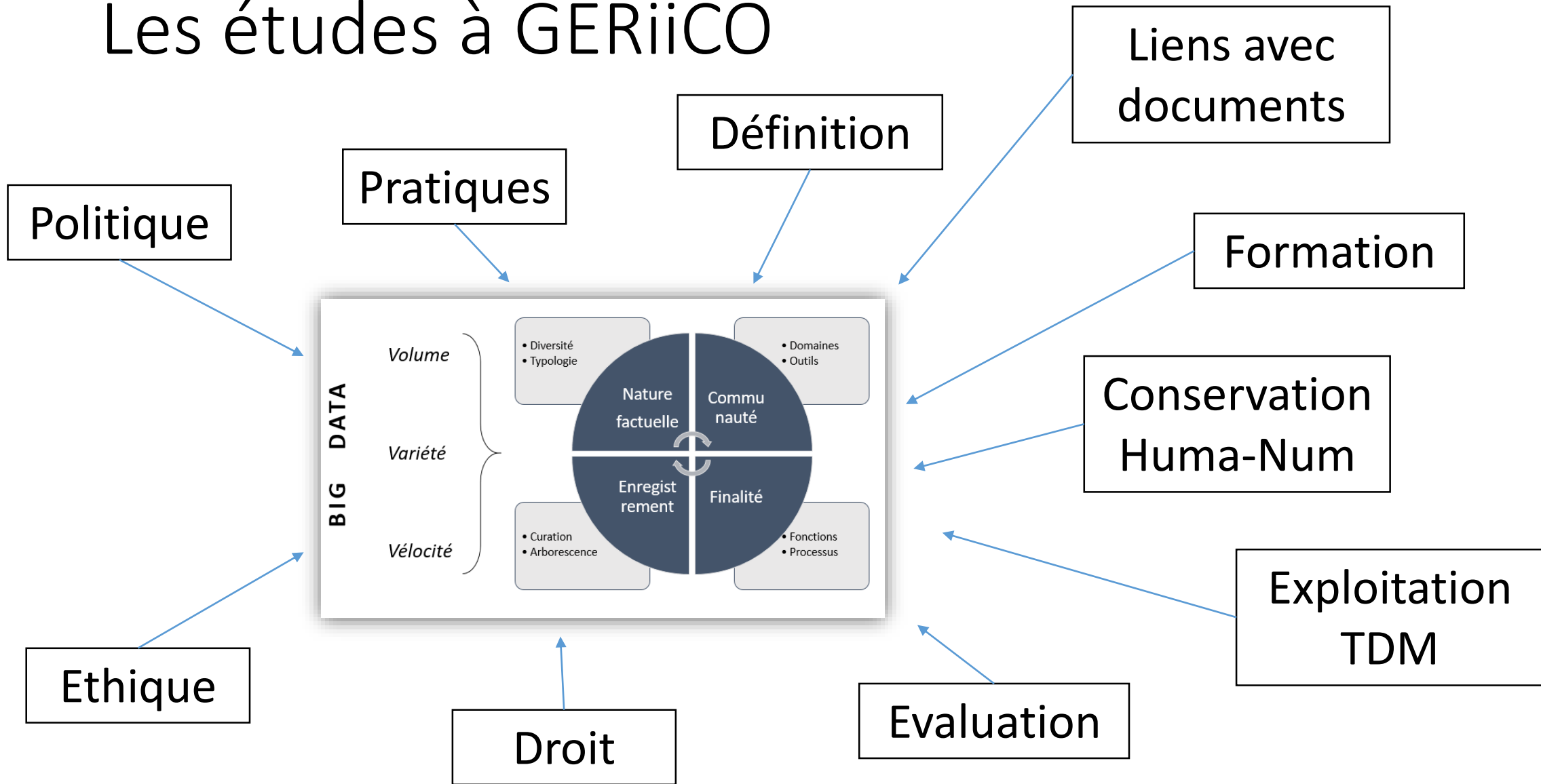
Finalité

- Curation
- Arborescence

- Fonctions
- Processus



Les études à GERiICO

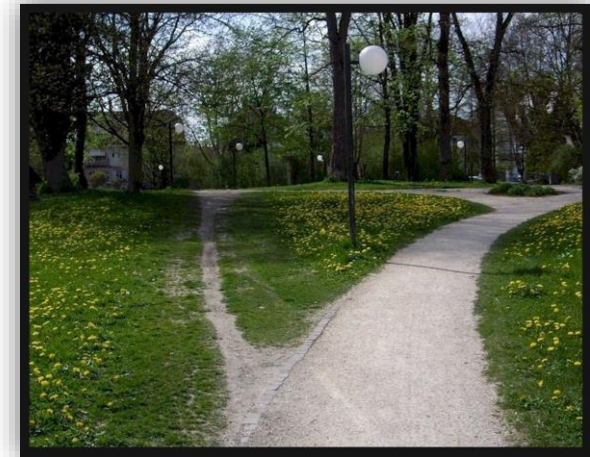


Interrogation normative

- Pourquoi : importance des « standards » dans les documents et discours sur les données de la recherche
- En même temps, ressentie d'une grande nouveauté, « hors norme »
- Définition COSSI 2018 :
 - *« La norme est établie par un organisme de normalisation reconnu dans le cadre d'un processus (...) »*
 - *Le standard est un produit industriel qui se répand et tend à faire autorité (...)*
 - *L'une et l'autre procèdent de la recherche d'un acte optimisé, d'un one best way de la qualité, de la production, de l'usage...*
 - *L'une et l'autre font l'objet de stratégies industrielles et d'influence dans l'établissement de la norme aussi bien que du standard. »*

Pour la suite, une approche plus large

- Référentiels incontestables communs
 - lois
 - directives
 - normes
 - recommandations
 - bonnes pratiques
- Ce qui doit être
- Etat habituel / standard *de facto*
- Conforme à la majorité des cas

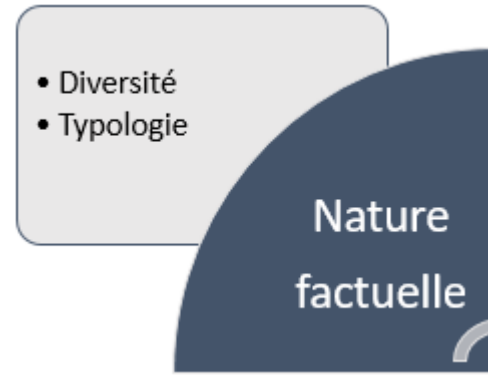


L'environnement normatif des données de la recherche

A partir des quatre dimensions du concept des données de la recherche

(1) Nature factuelle

- Terminologie
 - Définitions
 - Données primaires vs données secondaires
 - Répertoires
 - re3data
 - Cat-OPIDoR
- Données personnelles
- Données de santé

- 
- Diversité
 - Typologie

Nature
factuelle

L'approche typologique

Exemple d'une classification de données (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.

Répertoire *re3data*

13 catégories

Une répartition très inégale

Plusieurs larges catégories,
transversales aux disciplines, aux
contours mal définis

Une catégorie « autres » : une longue
traîne d'autres types de données
dans 1/3 des entrepôts de données

	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)

Données personnelles

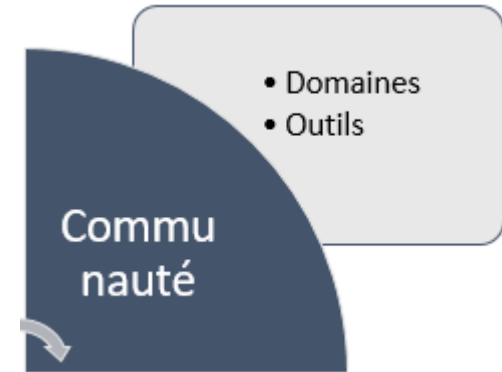
- Loi Informatique et libertés
 - Obligations, droits
 - Nature des données, traitement des données, finalité
 - Sécurité du SI
- Règlement général sur la protection des données (RGPD)
 - Données de la santé

RGPD : se préparer en 6 étapes



Le 25 mai 2018, le règlement européen sera applicable. De nombreuses formalités auprès de la CNIL vont disparaître. En contrepartie, la responsabilité des organismes sera renforcée. Ils devront en effet assurer une protection optimale des données à chaque instant et être en mesure de la démontrer en documentant leur conformité.

(2) Communauté



- Standards disciplinaires
 - Recommandations, bonnes pratiques (archéologie, santé...)
 - Procédures de production (protocoles etc.)
 - Publication des données (cf. APA)
- Equipements avec leurs propres normes (observatoire, IRM etc.)
- Méthodologies (enquêtes, observations, statistiques etc.)
- Définition des données (cf. web analytics)
 - Profil disciplinaire des données (typologie, volume, mode de production...)
- Vers un référentiel de bonnes pratiques
 - partage, transparence, répliquabilité

L'aspect éthique des données en SHS

Cf. Jacquemin et al. (2018)

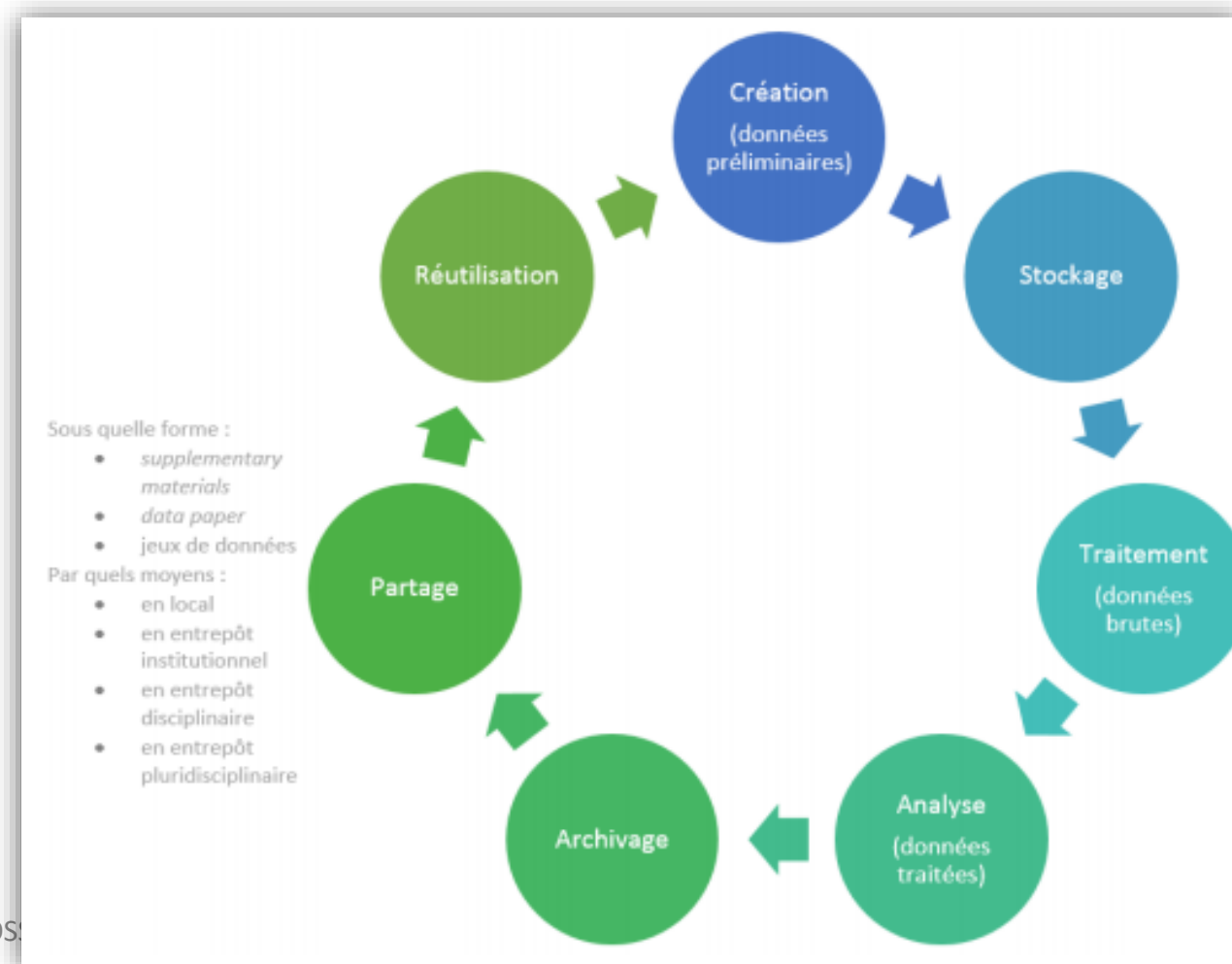
- Protocole éthique et plan de gestion des données
 - H2020 *ethics reviews*
- Données personnelles
- Respect des personnes, conflits d'intérêt
- Crédibilité des données
 - Certification des entrepôts (cf. *Core Trust Seal*)
- Sécurité des données
- Propriété intellectuelle

(3) Finalité

Finalité

- Fonctions
- Processus

- 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



Approche fonctionnelle

- Politique
 - Loi numérique, Plan d'action etc.
 - Données scientifiques et données publiques
 - 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, l'aménagement du territoire..
- Economique
 - Interopérabilité, ouverture (principes FAIR)
- Scientifique
 - Lien avec l'administration et le financement de la recherche
 - H2020, ANR, HCERES etc.

Cycle de vie des données

- Référentiels de la production des données
- Normes de sécurité pour l'archivage numérique
- Normes et standards de la communication des données
 - COUNTER/SUSHI pour données d'usage des ressources en ligne
 - Protocole d'échange de données NF ISO 20614 (cadre de transactions pour données et métadonnées)
- Normes de l'archivage (NF, ISO...)
- Recommandations pour le traitement des données
 - Signification statistique
- Incitations au partage

(4) Enregistrement

- Normalisation des métadonnées
 - Génériques vs spécifiques
 - Dublin Core
 - DataCite Metadata Schema
 - Data Documentation Initiative (DDI)
- Normalisation des identifiants
 - DOI (ISO)
 - ORCID (compatible avec ISNI/ISO)
 - OrgID (à venir)
 - Nomenclature of Celestial Objects (Centre de Données astronomiques de Strasbourg)

Enregist
rement

- Curation
- Arborescence

DataCite Metadata Schema 4.1 (2017)

<https://schema.datacite.org/>

- 19 champs
 - dont 6 champs obligatoires
 - et 13 champs recommandés ou optionnels
- compatibles avec DC

<i>ID</i>	<i>Property</i>	<i>Obligation</i>
1	Identifier (with mandatory type sub-property)	M
2	Creator (with optional family name, given name, name identifier and affiliation sub-properties)	M
3	Title (with optional type sub-properties)	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType (with mandatory general type description sub-property)	M

La force des normes

Ou : Qu'est-ce qui fait la force des normes - la sanction, la contrainte, leur caractère obligatoire ?

Les plans de gestion

- Cahier des charges pour les projets du programme H2020
 - Avec l'application des principes FAIR
- Incitation par plusieurs établissements
 - Avec modèles, conseils et formations
- Engagement du Plan d'Action National 2018-2020
- Future exigence par l'ANR ?

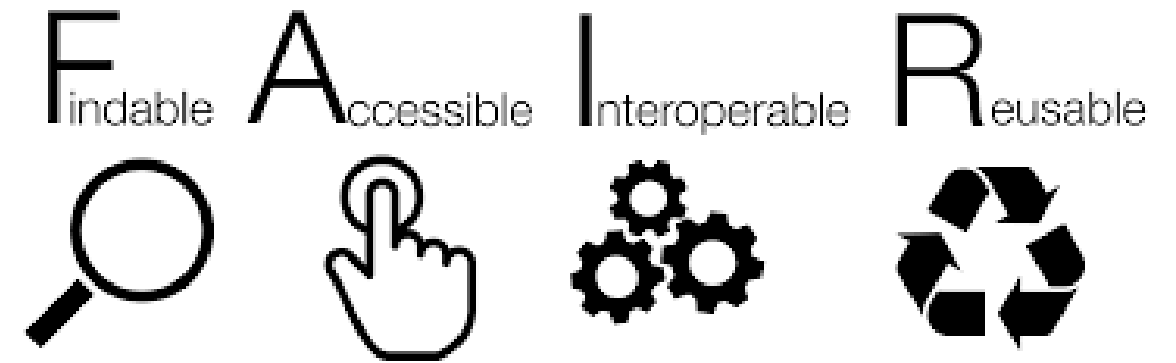
Protocoles éthiques

- Politique scientifique des organismes
 - COMETS du CNRS
- Etablissements
 - Comités d'éthique, consignes, bonnes pratiques
- Programmes de recherche
 - H2020, ANR etc.
- Communautés scientifiques
 - Cf recherche biomédicale (Code Nuremberg, déclarations de référence...)
- Législation
 - Données personnelles
 - Recherche biomédicale (santé publique, bioéthique...)
- Directives européennes

Evaluation

Cf. Schöpfel et al. (2016)

- Prise en compte des données de la recherche par les systèmes d'évaluation
- Evaluation de la gestion, pas de qualité des données
 - Autrement dit, évaluation des principes FAIR
- Format standard CERIF (euroCRIS)
- Autres formats industriels, plus ou moins interopérables
 - Elsevier avec PURE
 - Caplab ? Conditor ?



La gestion FAIR

Cf. INRA, Principes FAIR

<https://www6.inra.fr/datapartage/Technologies/Principes-FAIR>

Des recommandations techniques

- Pour le développement des infrastructures « ouvertes »
 - machine readable
- Pas pour le fonctionnement « humain »
- Deviennent des standards *de facto*
 - *H2020, EUDAT...*
- Cadre de référence pour une stratégie « données »
 - Par extension, aussi pour la diffusion des documents
 - Et pour la définition d'une politique de science ouverte
- En fait, une cascade de normes

Findable

- Identifiants standards
 - DOI, handle, URI
- Métadonnées riches
 - standards données (DataCite)
 - standards disciplinaires
- Mécanismes d'interrogation standards
 - API
 - SPARQL, SQL

Accessible

- Accessibilité via un protocole d'accès standardisé
 - Http
 - API REST
- Dépôt dans entrepôt certifié
 - accès ouvert
 - certification
- Métadonnées disciplinaires standardisées

Interoperable

- Utilisation d'un langage formel pour la représentation des connaissances
 - Web Ontology Language (OWL)
 - Resource Description Framework (RDF)
 - Simple Knowledge Organization System (SKOS)
 - etc.
- Terminologies normalisées (largement partagées)
- Standards disciplinaires

Reusable

- Mise à disposition selon une licence explicite et accessible
 - Creative Commons, autres licences ouvertes
- Formats et métadonnées standards
- Indication claire de la provenance des données

Conclusion

On aurait pu procéder autrement...

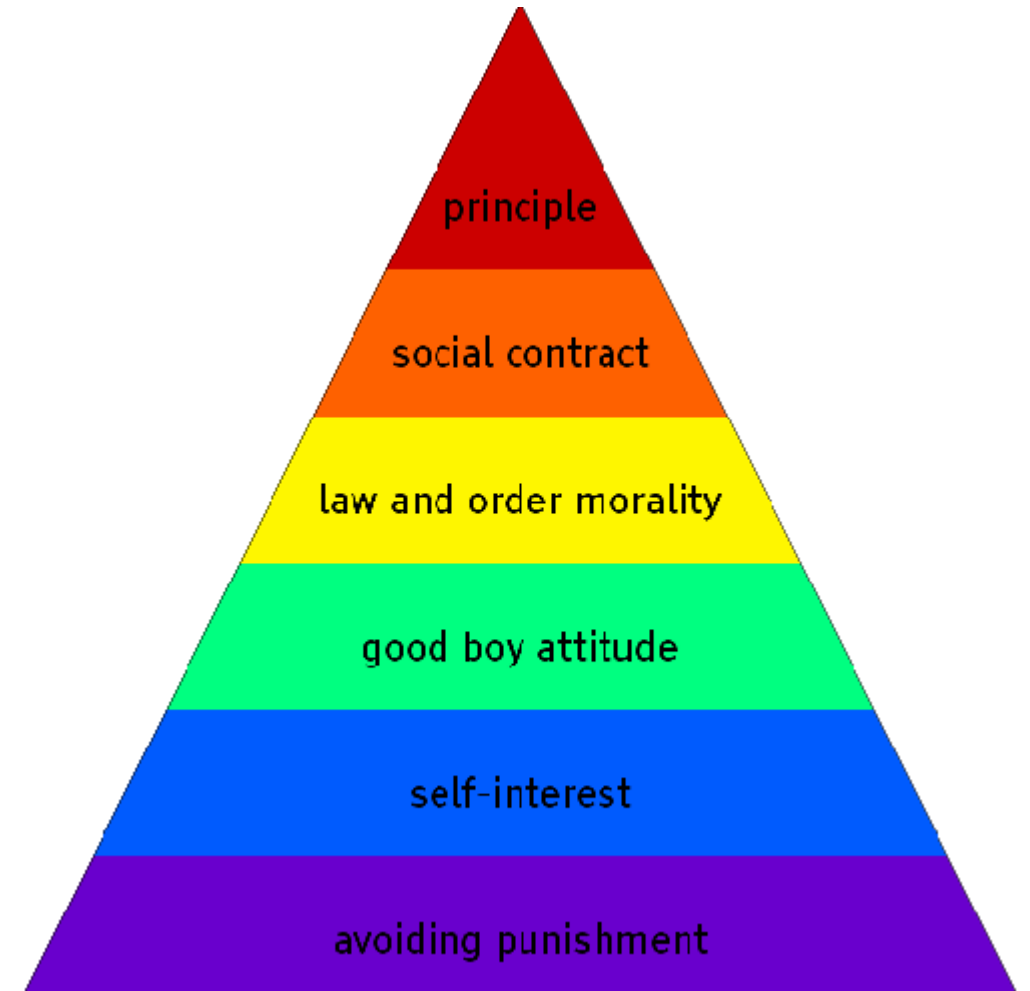
- Analyser les études et rapports sur la gestion des données
 - CNRS DIST, COPIST, DataCite, Commission Européenne
- Présenter les projets de DataCite
 - dont un « Code of Practice » pour l'attribution des DOI, à l'instar du standard COUNTER
- Décrire l'approche normative de la [Research Data Alliance](#)
 - NISO/Privacy
 - Common standards for DMP
 - Legal interoperability
 - etc.

Ou bien....

- Distinguer au moins trois types de normes
 - normes légales
 - normes techniques
 - normes éthiques
- Classer les normes selon leur degré de contrainte
 - lois
 - normes
 - contrats
 - recommandations
 - bonnes pratiques
- Parler des acteurs, dont les labos, SCD, DSI , Direction Recherche etc.
- Parler aussi de la force normative des faits

(1) Du point de vue psychologique (Kohlberg)

- Comment fonctionnent les communautés scientifiques ?
- Quelle est la stratégie des comités, organismes, établissements, services ?
- L'impact des bonnes pratiques disciplinaires ?
- L'intérêt d'un « modèle Liège » pour les données ?



(2) L'expérience utilisateur (SHS)

- La gestion des données comme pratique artisanale
 - Sur mesure, particulier, hétérogène, qualité...
- Le partage des résultats comme pratique normale
- Connaissance partielle des normes
 - Les référentiels du domaine d'expertise
 - Normes ressenties comme contraintes surtout en absence de solution
- La politique d'ouverture des données comme contrainte externe
 - Finalité politique et industrielle vs objectifs personnels
- Légitimité et crédibilité des services et métiers
 - SCD, DSI, Direction Recherche, comité d'éthique, laboratoires...

(2a) Deux logiques

- Travail scientifique
- Gestion de production



Avec une exploitation industrielle



(3) Questions ouvertes

- Normes vs ressources
 - Une politique sans moyens ?
- Normes pour qui ?
 - *Cui bono* ?
 - Quid de la recherche « privée » ?
- Injonctions contradictoires ?
 - « as open as possible, as closed as needed »
 - ouverture publique, fermeture privée ?
 - ouverture vs valorisation ?
- Priorités des besoins et urgences ?
- Hiérarchie des sanctions ?

Références

- André, F., 2015. Déluge des données de la recherche ? In: Calderan, L. et al. (dir.), Big data : nouvelles partitions de l'information. Actes du Séminaire IST Inria, octobre 2014. De Boeck; ADBS, Louvain-la-Neuve, pp. 77-95.
- Borgman, C. L., 2016. Big data, little data, no data : scholarship in the networked world. The MIT Press, Cambridge MA.
- Jacquemin, B. et al., 2018. L'éthique des données de la recherche en SHS. In: DocSoc2018, 6e conférence "Document numérique & Société", Echirolles, 27 et 28 septembre 2018.
- Schöpfel, J. et al., 2016. Research data in current research information systems. In: CRIS 2016, St Andrews, 8-11 June 2016.
- Schöpfel, J. et al., 2017. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. In: Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017, 31 mai 2017 Toulouse (France).
- Serres, A. et al., 2017. Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2.

Merci

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

Contact : joachim.schopfel@univ-lille.fr