Les données de la recherche

Politiques et enjeux

Joachim Schöpfel

Plan

- Introduction
  - Le cadre du projet
  - Le concept des données de la recherche
- Trois niveaux politiques (« bottom-up »)
  - Le terrain des unités de recherche et établissements
  - Le périmètre national
  - L'espace européen de la recherche
- Tensions
- Et le chercheur ?

Introduction

« La politique (...) se réfère à la pratique du pouvoir » (Wikipédia)

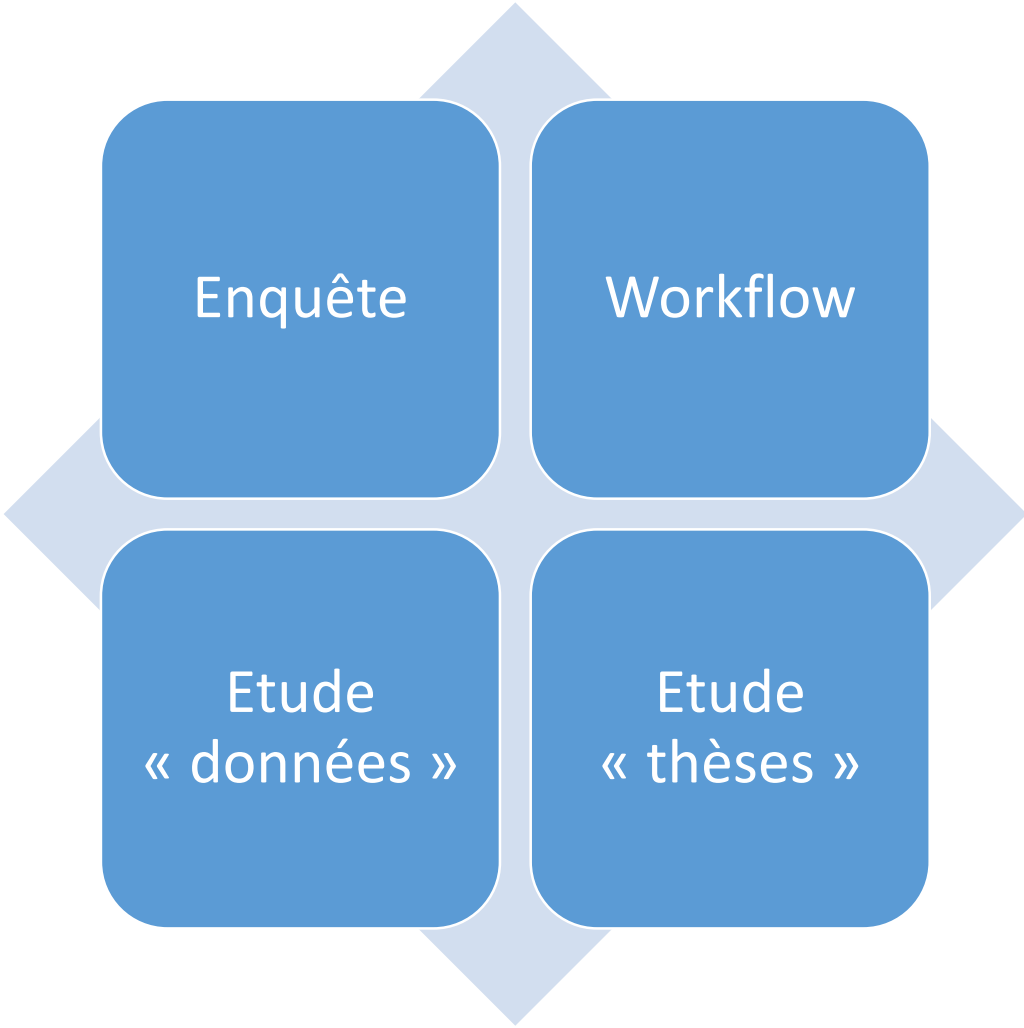
Le projet *D4Humanities*

Projet structurant 2017-2018

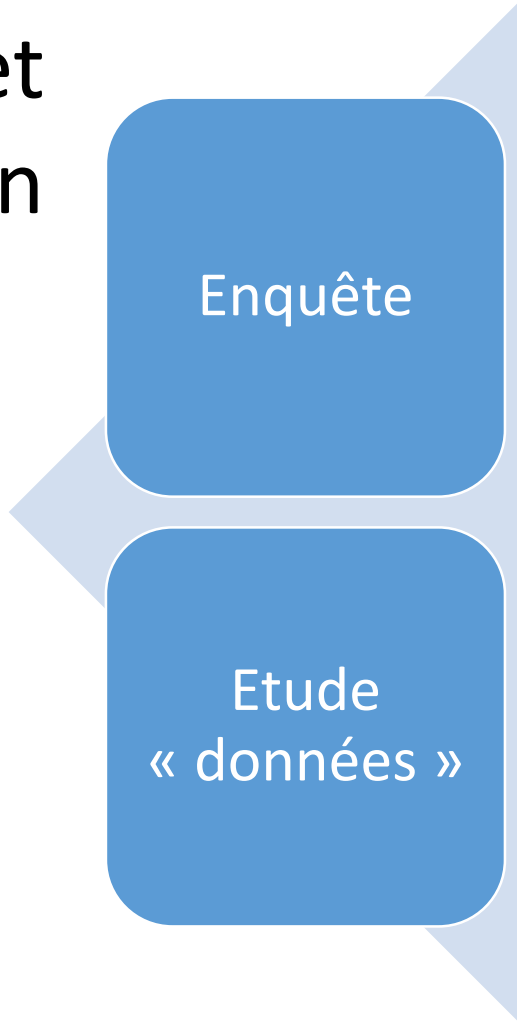
Participants : plusieurs laboratoires, SCD Lille SHS, MESHS, Ecole Doctorale SHS, ANRT ; suivi par la Direction Recherche

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

Suite : préparation d'un projet franco-allemand ANR/DFG 2019 (*xDiss*)



Etude des pratiques et besoins sur le terrain



- Concept
- Normes
- Ethique
- Politiques
- (Typologie)

L’approche conceptuelle

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

L’obscur objet…

« (…) des enregistrements factuels (chiffres, textes, images et sons), utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche » (OCDE)

« Tout est donnée » (Chignard)

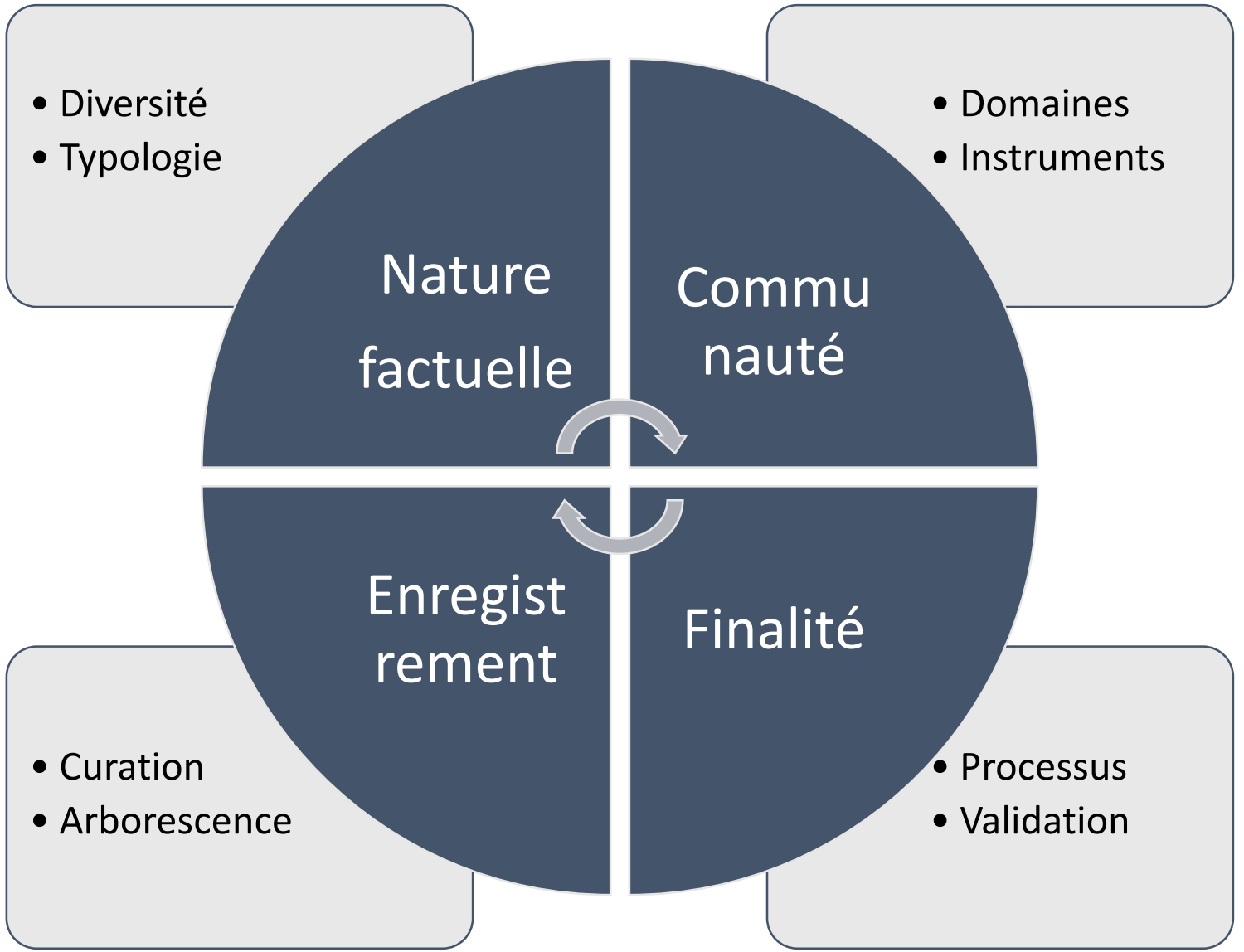
« Data is fuel of economy » (Kroes)

BIG DATA

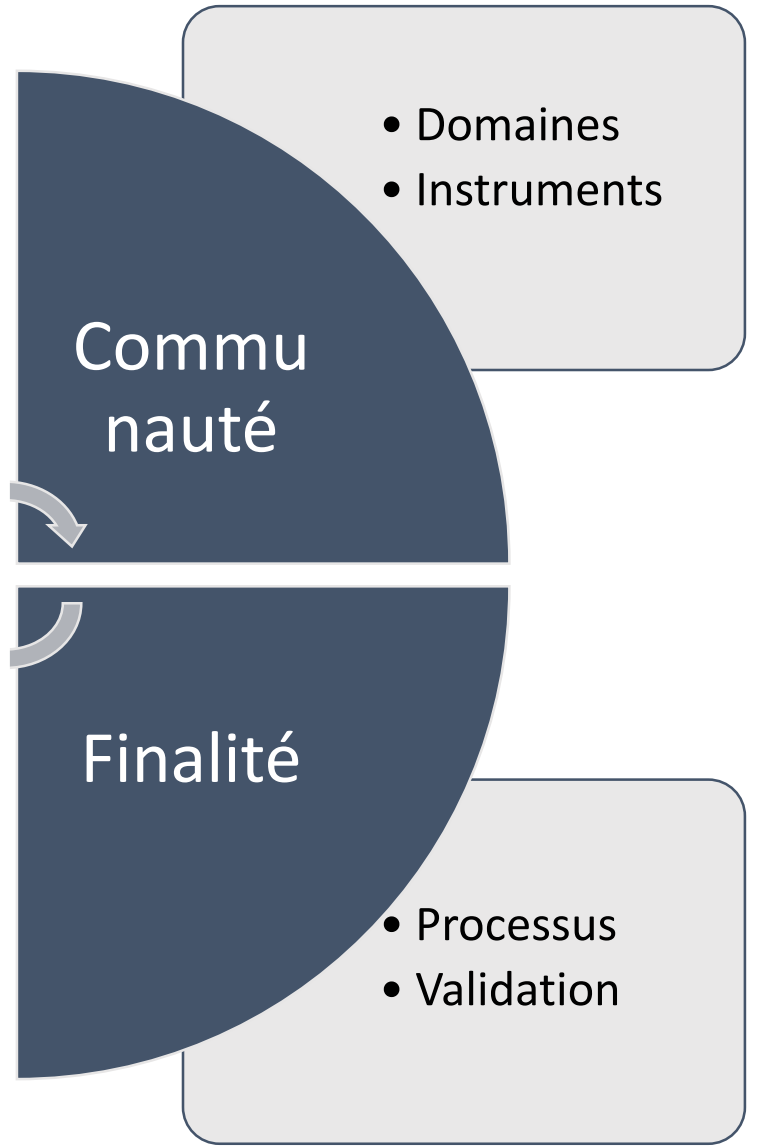
Volume

Variété

Vélocité



BIG DATA


• Domaines
• Instruments
Commu
nauté
Finalité
• Processus
• Validation

Approche fonctionnelle

Politique

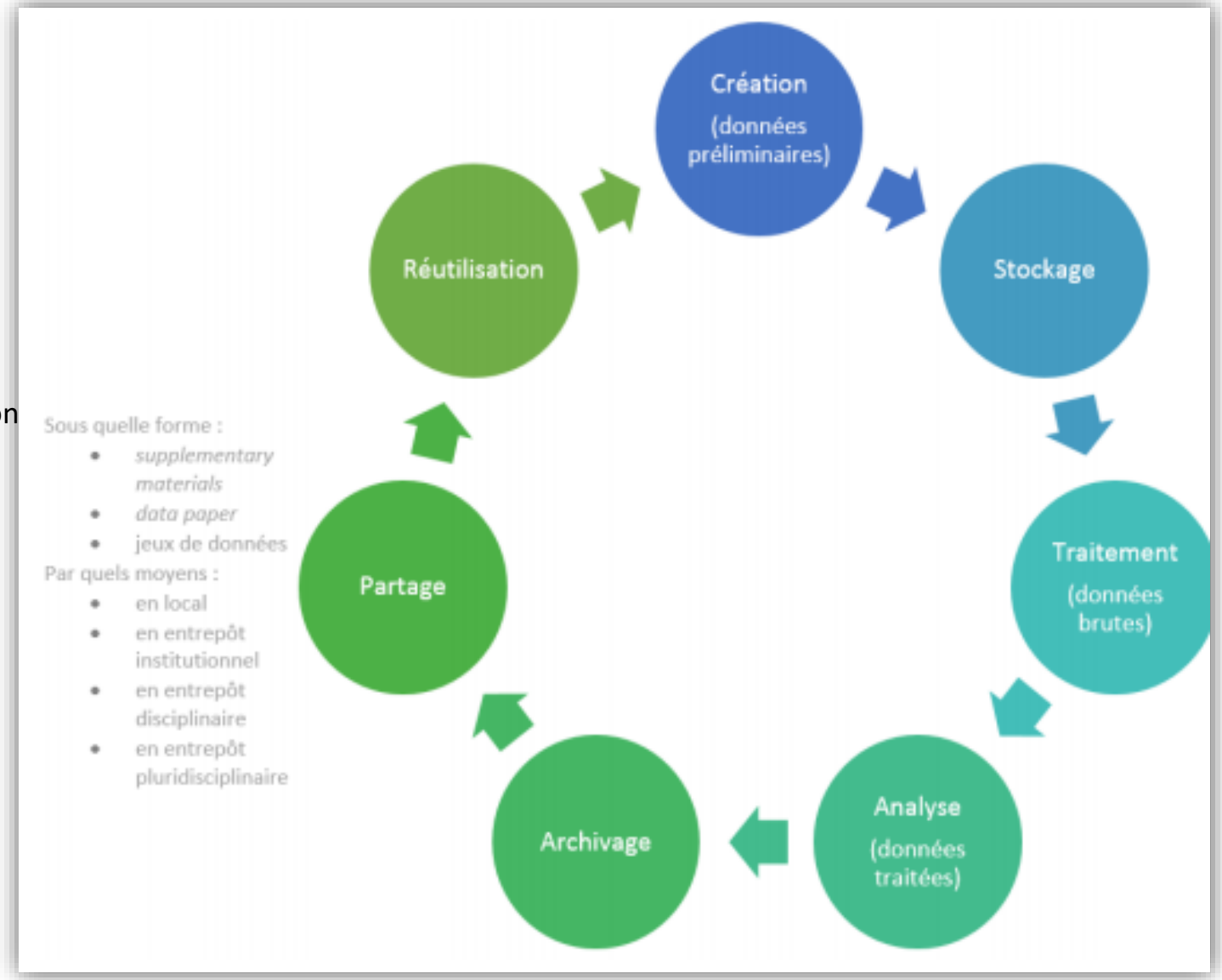
- Augmenter la transparence
- Créer un environnement favorable à l'innovation
- Rendre l'action publique plus efficace

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)

Trois niveaux politiques (« bottom-up »)

« La politique est l’art du possible » (Gambetta)

(1) Contexte politique – le terrain

- Enjeux
  - bons outils de travail
  - pratiques efficaces
  - conformité avec réglementations
  - conformité avec conditions des agences et éditeurs
  - (ranking/évaluation)
- Métiers
  - chercheurs
  - ingénieurs et techniciens de laboratoires
  - informaticiens
  - juristes
  - bibliothécaires, documentalistes
- Structures
  - universités
  - organismes des recherche
  - laboratoires

Enquête sur le campus de Lille SHS 2017-2018

- La sécurité des données et des systèmes
- Une communication compliquée
- Un continuum de pratiques
- Disciplines ou méthodes ?
- **Plusieurs niveaux de gouvernance**
- **Incitations**
- **Verrous**

Plusieurs niveaux de gouvernance

- Qui est responsable des données de la recherche ? Qui coordonne leur gestion ? Qui a une vue globale de la collecte et production des données ?
- Les données sont souvent considérées comme une affaire personnelle, voire comme une propriété privée, sous la responsabilité du chercheur
- Une gouvernance à plusieurs niveaux – parfois structurée et explicite, mais souvent diffuse et plus ou moins informelle
- Projet vs laboratoire (vs établissement)

Incitations

- Quels facteurs favorisent les bonnes pratiques ?
  - Programmes scientifiques
    - H2020, demain ANR (cf. Plan d'action national 2018-2020)
    - Plans de gestion
  - Les revues
    - APA, Nature etc.
  - Législation
    - Recherche, santé publique, données personnelles, open data
- Comité d'éthique
- Partenaires

Verrous

- Manque de ressources
  - Informatique
  - Ressources humaines
    - Ingénieurs de données, éthique, DSI etc.
  - Moyens financiers
- Manque de motivation ?
- Ecosystème « traditionnel » ?

Propositions pour une politique de données

- **Mettre en place un pilotage scientifique**
- Investir d'une manière ciblée
- **Viser les projets, pas les laboratoires**
- Utiliser les plans de gestion comme levier
- Apporter des réponses aux contraintes de sécurité
- Apporter des réponses aux besoins de communication
- Apporter des réponses aux besoins de curation
- **Proposer plusieurs solutions pour la conservation des données**
- **Institutionnaliser le lien avec le TGIR Huma-Num**
- Soutenir les bonnes pratiques

Mettre en place un pilotage scientifique

- Un pilotage scientifique par un comité de pilotage et de coordination rattaché à la Direction de la Recherche
- Avec des compétences politiques et scientifiques
- Pour la préparation d'une politique de données à décider par les conseils centraux, et la coordination de sa mise en œuvre
  - Réunissant le Vice-Président Recherche, des représentants des laboratoires et projets scientifiques, des représentants des Directions Recherche (valorisation, ingénierie et management de projets), Système d'information et Affaires juridiques, des responsables SSI et RGPD, du président du comité éthique et du SCD

Viser les projets, pas les laboratoires

- Concentrer le développement d'une offre de service aux projets
  - Les laboratoires comme vecteur de communication
  - Les projets comme lieu de conseil, d'assistance, de formation, d'accompagnement
    - Phase de montage
    - Suivi
    - Conservation et valorisation des résultats
- L'intérêt d'une telle approche
  - *Des besoins précis et immédiats.*
  - *Des contraintes imposées par le financement.*
  - *Des obligations légales et réglementaires.*
  - *Une gouvernance plus simple.*
  - *L'expérience des pratiques collaboratives au sein des équipes.*

Proposer plusieurs solutions pour la conservation des données

- Un dispositif modulaire
  - Sur le campus
  - En dehors du campus
- Pour les données « chaudes »
- Pour les données « froides »
  - Réglementations spécifiques
  - Durées variables
- Les infrastructures nationales
  - CINES, Huma-Num, RENATER etc.
  - Demain, un DataVerse français ?

Institutionnaliser le lien avec le TGIR Huma-Num

- Désignation d'un correspondant local TGIR Huma-Num
  - à l'instar de l'Université de Nice Sophia Antipolis
- Pour promouvoir le dispositif et coordonner et faciliter les contacts avec les SHS de Lille
- Pour renforcer les liens entre SCD et MESHS
- Pour contribuer au développement de l'offre de service du TGIR
- Pour développer la présence de l'Université de Lille dans les infrastructures européennes DARIAH et CLARIN dans lesquelles Huma-Num représente la France

Pilotage, services, recherche, enseignement

- Pilotage politique
- Fédération des services
- Programme de recherche autour de la science ouverte
- Programme d’enseignement

(2) Contexte politique – France

- Code de la Recherche, Article L112-1
  - La recherche publique a pour objectifs : (...) e) L'organisation de l'accès libre aux données scientifiques.
- Loi numérique 2016
  - Droit de l'exploitation secondaire, open data, TDM
- Pour une action publique transparente et collaborative : plan d'action national pour la France 2018-2020, Engagement 18
  - https://www.etalab.gouv.fr/wp-content/uploads/2018/04/PlanOGP-FR-2018-2020-VF-FR.pdf
- Plan national pour la science ouverte
  - http://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html

Plan d’action national 2018-2020 (avril 2018)

- Communiquer auprès des communautés scientifiques sur les implications de la loi numérique relatives à l'ouverture des publications et des données (2018 ou 2019)
- Dans le cadre du soutien public aux revues, recommander l'adoption d'une politique de données ouvertes associées aux articles et le développement des data papers (n.d.)
- 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 (2019 et en continu)

https://www.etalab.gouv.fr/wp-content/uploads/2018/04/PlanOGP-FR-2018-2020-VF-FR.pdf

Plan national (juillet 2018)

- Trois axes prioritaires
  - Axe 1 Généraliser l'accès ouvert aux publications
  - Axe 2 Structurer et ouvrir les données de la recherche
  - Axe 3 S'inscrire dans une dynamique durable, européenne et internationale
- Neuf mesures
- Dix domaines d'action
- 27 actions

Généraliser l'accès ouvert aux publications

L'ouverture des publications scientifiques doit devenir la pratique par défaut aussi vite que possible. Pour engager cette dynamique, les publications issues de recherches financées au moyen d'appels à projets sur fonds publics seront obligatoirement mises à disposition en accès ouvert, que ce soit par la publication dans des revues ou ouvrages nativement en accès ouvert, soit par dépôt dans une archive ouverte publique comme HAL.

" La recherche scientifique est un bien commun que nous devons partager avec tous. "

Les mesures

1. Rendre obligatoire la publication en accès ouvert des articles et livres issus de recherches financées par appel d'offres sur fonds publics.

2. Créer un fond pour la science ouverte.

3. Soutenir l'archive ouverte nationale HAL et simplifier le dépôt par les chercheurs qui publient en accès ouvert sur d'autres plateformes dans le monde.

Les actions

Reconnaître la science ouverte

- Reconnaître la science ouverte dans les évaluations des chercheurs et des établissements.
- Réduire l'emprise de l'évaluation quantitative au profit de l'évaluation qualitative.
- Encourager l'adoption des citations ouvertes (Initiative for Open Citations – I4OC) à la place de citations dans des environnements propriétaires.

Construire la bibliodiversité

- Explorer les nouveaux modèles économiques pour les revues comme pour les livres en accès ouvert.
- Dynamiser nos presses universitaires et notre secteur éditorial qui feront le choix de l'accès ouvert.
- En cas de frais de publication, les réserver aux publications entièrement en accès ouvert.

Piloter la science ouverte

- Mettre en place un baromètre de la science ouverte en France

Structurer et ouvrir les données de la recherche

Notre ambition est de faire en sorte que les données produites par la recherche publique française soient progressivement structurées en conformité avec **les principes FAIR** (Facile à trouver, Accessible, Interopérable, Réutilisable), préservées et, quand cela est possible, ouvertes.

> *Les données de la recherche sont la matière première de la connaissance. Les partager, c'est ouvrir de nouvelles perspectives scientifiques .*

Les mesures

4 Rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics.

5 Créer la fonction d'administrateur des données et le réseau associé au sein des établissements.

6 Créer les conditions et promouvoir l'adoption d'une politique de données ouvertes associées aux articles publiés par les chercheurs.

Les actions

Accélérer

- Proposer un appel ANR Flash destiné à accélérer l'adoption des principes FAIR et l'ouverture des données de la recherche en France.
- Créer un prix des données de la recherche récompensant les équipes et projets exemplaires dans ce domaine.

Coordonner

- Construire autour de l'administrateur des données un réseau de correspondants dans les établissements, pour répondre aux questions que se posent les chercheurs sur les données de la recherche.
- Dans le cadre du soutien public aux revues, recommander l'adoption d'une politique de données ouvertes associées aux articles, le développement des articles de données et des revues de données.

Structurer

- Généraliser la mise en place de plans de gestion des données dans les appels à projets de recherche
- Développer des centres de données thématiques et disciplinaires.
- Développer un service générique d'accueil et de diffusion des données simples.
- Engager un processus de certification des infrastructures de données.

Organiser

- Soutenir la ***Research data alliance*** (RDA) et créer le chapitre français de l'alliance (RDA France).
- Soutenir ***Software heritage***, la bibliothèque des codes sources

S’inscrire dans une dynamique durable, européenne et internationale

Le succès de la science ouverte implique le développement de nouvelles pratiques quotidiennes pour les chercheurs. Cela nécessite la définition de nouvelles compétences, le développement de nouvelles formations et l'adoption de nouveaux services. Le Comité pour la science ouverte, qui rassemble plus de 200 experts du domaine, travaillera à la définition des nouvelles compétences nécessaires.

" La France s’engage pour que la science ouverte devienne la pratique quotidienne par défaut des chercheurs. "

Les mesures

7 Développer les compétences en matière de science ouverte notamment au sein des écoles doctorales.

8 Engager les opérateurs de la recherche à se doter d'une politique de science ouverte.

9 Contribuer activement à la structuration européenne au sein du *European Open Science Cloud* et par la participation à *GO FAIR*.

Les actions

Généraliser les compétences de la science ouverte

- Communiquer auprès des communautés scientifiques sur les implications de la loi numérique relatives à l'ouverture des publications et des données.
- Créer un label « science ouverte » pour les écoles doctorales.
- Développer les compétences sur les données de la recherche, notamment à travers des offres de formation en ligne à destination de la communauté scientifique.

Participer à l'échelle européenne et internationale au paysage de la science ouverte

- Créer un Comité pour la science ouverte regroupant les experts du domaine et qui traitera des publications, des données de la recherche, des compétences et de l'articulation avec l'Europe et l'échelle internationale. Il sera chargé de proposer une mise à jour du plan dans deux ans.
- Adhérer au niveau national à ORCID, système d'identification unique des chercheurs qui permet de connaitre plus simplement et sûrement les contributions scientifiques d'un chercheur.
- Créer la Fondation franco-néerlandaise DOAB (*Directory of open access books*).
- Contribuer aux infrastructures de la science ouverte comme le DOAJ, OpenAIRE, SCOSS, OPERAS, Crossref et DataCite.
- Coordonner les négociations avec les éditeurs à l'échelle internationale.

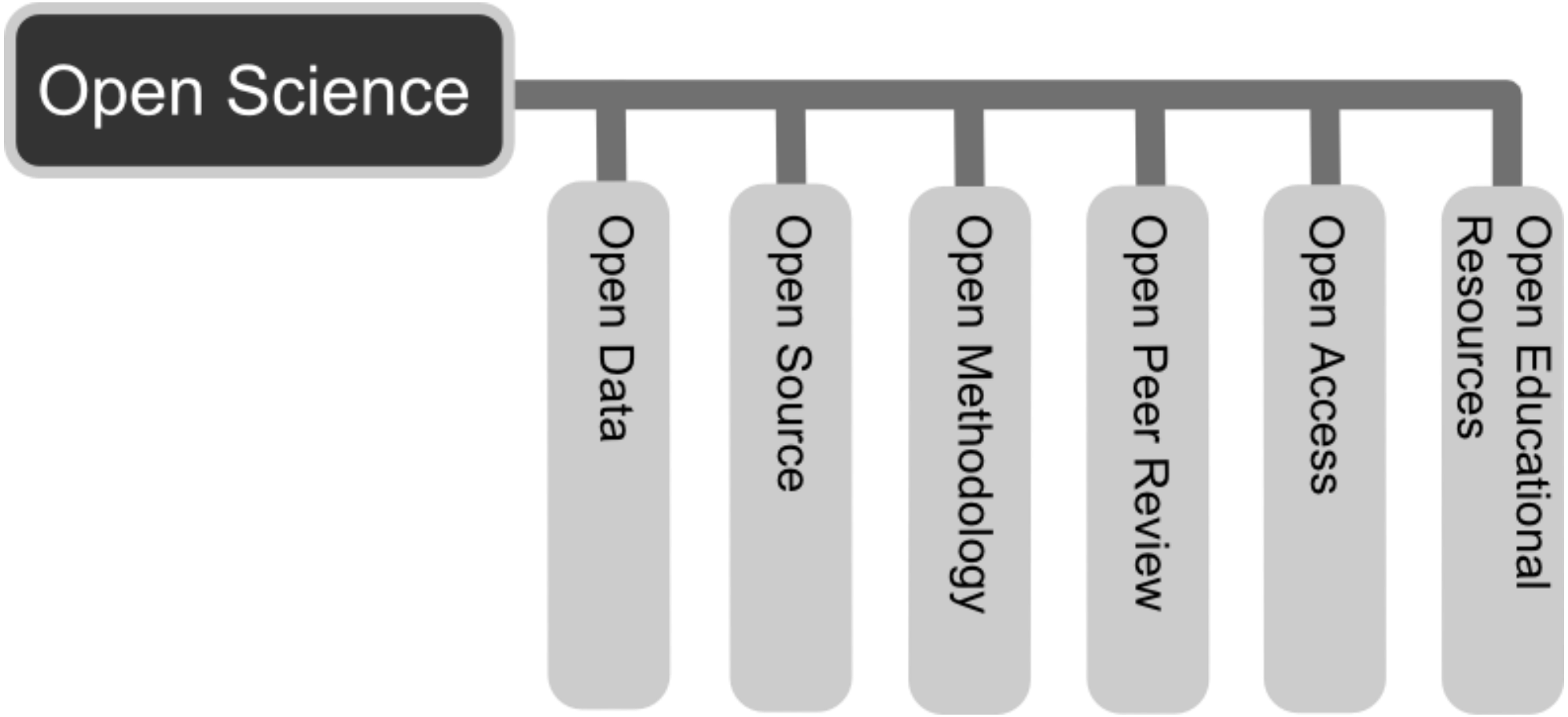
Participer à la transparence dans le cadre de l'open government partnership (OGP)

- Ouvrir les données du financement de la recherche en constituant des jeux de données publiques concernant
  - les dépenses relatives aux acquisitions électroniques dans les bibliothèques universitaires et par les organismes de recherche,
  - les dépenses relatives aux frais de publications d'articles et de livres
  - les financements de recherche sur appel à projets et leurs bénéficiaires.
- Enrichir scanR, moteur de la recherche et de l'innovation et Isidore, plateforme de recherche permettant l'accès aux données numériques des sciences humaines et sociales (SHS), et développer leur notoriété ainsi que leur usage afin d'alimenter le débat public autour des résultats de la recherche.

(3) Contexte politique – Europe

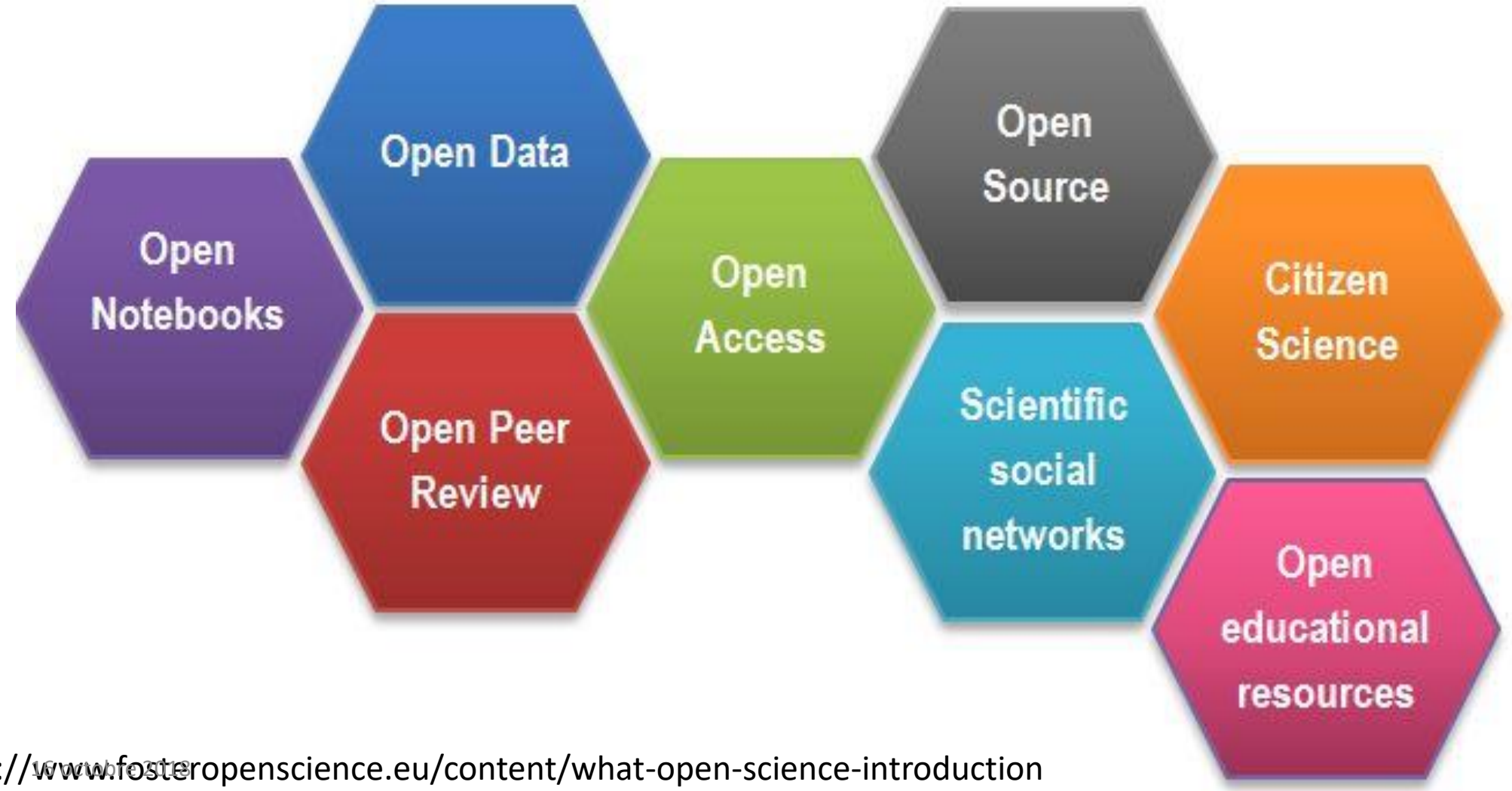
- Conseil européen *Call for action* Amsterdam 2016
  - Ouverture des publications et données
  - *As open as possible, as closed as necessary*
  - Ecosystème (= avec l'ensemble des acteurs)
- Communiqué G7 septembre 2017
  - *G7 Science Ministers committed to giving incentives for open science and to providing research infrastructures on the basis of FAIR data*
- European Open Science Cloud (EOSC) 2018
- Premier *GO FAIR Meeting* en France mars 2018
  - *The rationale for prioritizing machine readable metadata over simple text*

Que veut dire « science ouverte » ?



https://en.wikipedia.org/wiki/Open_science

Un concept ouvert, évolutif



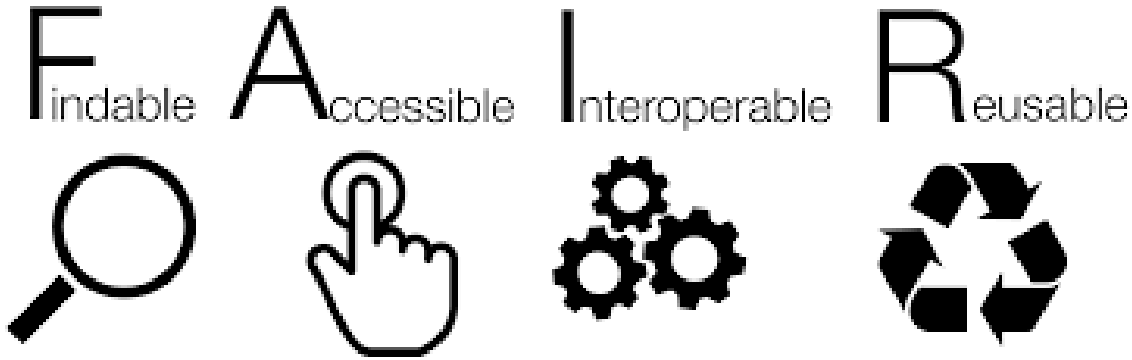
https://www.fosteropenscience.eu/content/what-open-science-introduction

Programme H2020

- Principe du libre accès aux publications et données
  - Mais protection des résultats avant publication
- Opération pilote « Open Research Data » (pour une partie du programme)
  - Libre accès aux données et métadonnées des publications
  - Dépôt d'un plan de gestion (mises à jour)
  - Libre accès aux données si possible
- Pour les autres projets : incitation, option ORD

Trois aspects emblématiques

- Fuel for economy
  - *Open data* : diffusion des données publiques, y compris de la recherche
- As open as possible, as closed as necessary
  - Ouverture des résultats de la recherche publique
  - Protection de l'innovation
  - (Protection des données personnelles)
- FAIR principles
  - *Big data* : rendre les résultats de la recherche « machine-readable »
  - "To maximise the value of science"



La gestion FAIR

Cf. INRA, Principes FAIR

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

Des recommandations ~~techniques~~ politiques

- Pour le développement des infrastructures « ouvertes »
  - *machine readable*
- 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

GT "FAIR Data Maturity Model« (RDA 2018)

- Core criteria to assess the implementation level of the FAIR data principles
  - Contribute to growth and accelerate innovation in a global digital economy
  - Provide savings in money
  - Provide savings in time for researchers and organisations
  - Increase transparency
- Cf. l'ensemble des acteurs, initiatives, institutions dans ce secteur
  - Objectifs ? Financement ?
- https://rd-alliance.org/group/fair-data-maturity-model-wg/case-statement/fair-data-maturity-model-wg-case-statement

Tensions

« La politique est l'art d'empêcher les gens de se mêler de ce qui les regarde » (Paul Valéry)

Divergences

- Besoins du terrain vs intérêts économiques et industriels
- Protection de l'innovation vs partage des résultats de la recherche
- RGPD vs open data
- Manque de moyens (RH, équipements) vs restrictions budgétaires
- Secteur public vs secteur privé
- Appel de Jussieu vs OA2020 (cf. Romary 2018)
- Politique locale vs coordination nationale
- Positionnement des métiers et fonctions

Enjeux à venir

- Contrôle des données
  - L'industrie de l'information a toujours défendu l'idée du libre accès aux résultats de la recherche
- Contrôle de l'exploitation des données
  - Qui a la capacité d'investissement nécessaire ?
  - Demain, un Facebook de la recherche ?

Et les chercheurs ?

- « Si les données sont au cœur du travail des chercheurs, les chercheurs sont au cœur de toute politique des données (…) Or, aucune politique de description, de stockage, d'archivage, de partage des données ne pourra se mettre en place sans l'adhésion ou l'implication des chercheurs » (Serres et al. 2017)

Quelques références

- G. Chartron (2018). `L'Open science au prisme de la Commission européenne'. *Education et sociétés* 41(1):177-193.
- B. Jacquemin, 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.
- L. Romary (2018). `Open Access in France: how the call of Jussieu reflects our social, technical and political landscape'. In *Open-Access-Tage*, September 2018, Graz, Austria.
- J. Schöpfel (2018). `Hors norme ? Une approche normative des données de la recherche'. In *COSSI 2018, Processus de normalisation et durabilité de l'information*, Université de Bordeaux, 24 et 25 mai 2018.
- J. Schöpfel (2018). `*Vers une culture de la donnée en SHS. Une étude à l'Université de Lille*'. Université de Lille, Villeneuve d'Ascq.
- J. Schöpfel, 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).
- J. Schöpfel, et al. (2018). `Research Data Management in the French National Research Center (CNRS)'. *Data Technologies and Applications* 52(2):248-265.

Merci

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

Contact : joachim.schopfel@univ-lille.fr