Gestion des données de recherche au CNRS : stratégie institutionnelle et approches disciplinaires

Francis André
CNRS/DIST
francis.andre@cnrs-dir.fr

Gérer ?... mais gérer quoi ?

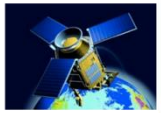
La principale difficulté que représente la définition, dans le champ de la recherche scientifique, de la notion de « donnée de recherche », est qu'il faut s'efforcer d'identifier ce qui peut rassembler des éléments aussi divers qu'un cliché de coléoptère prélevé à Madagascar, un spectrohéliogramme produit à Meudon, des informations sur les gènes d'une moisissure, les relevés météorologiques d'un vaisseau ayant traversé l'Atlantique au XVIIIe siècle ou l'enregistrement d'un dialecte rare. *(Rémi Gaillard, DCB/ENSSIB, 2014)*

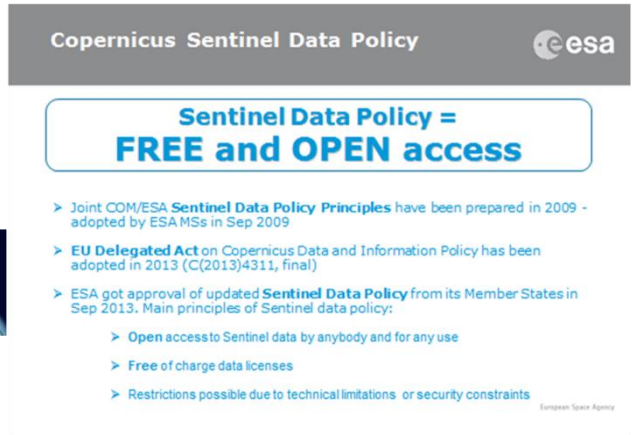
- Données primaires/secondaires
- Données expérimentales, d'observation, de simulation, dérivées, compilées, canoniques,
- Données brutes, élaborées, agrégées, enrichies, annotées, formatées, normalisées, traitées, publiées
- Données structurées/non structurées, homogènes/hétérogènes
- Données libres/protégées

La gestion des données…





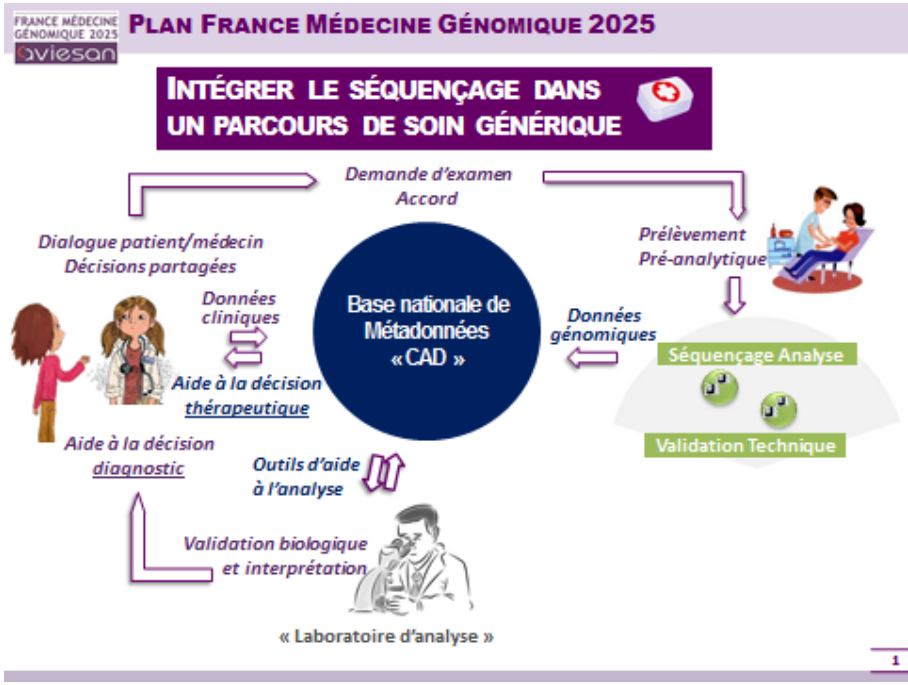


- Une organisation, une série d'instruments, une politique de données

- Toutes les disciplines sont concernées
- Big/small data
-

- Des barrières sociales, juridiques, comportementales,…



Données de la recherche ?

- Un objet plus que jamais numérique, dont l'importance dans le processus de construction de connaissances va croissante
- Toujours plus facile (trop ?) à produire en masse, de façon routinière
- Changement de paradigme : induction, déduction, abduction >> science des données
- La donnée s'enrichit au cours du processus de recherche et devient un objet à haute valeur ajoutée : de la donnée brute à la donnée publiée
- Dans une science plus ouverte, la donnée devient un élément partageable, accélérateur de l'innovation
- Examiner les principaux paramètres qui influent sur les rapports entre les chercheurs et leurs données
  - Contexte politique
  - Contexte scientifique
  - Infrastructures de données
  - Contexte juridique, protection des données personnelles
  - Paramètres disciplinaires, paramètres institutionnels
  - Cycle de vie de la donnée : curation
  - Autres : compétences, éthique, évaluation,…

OCDE 2004-2007

- Ministres de la Recherche et de la Technologies des pays de l'OCDE + Afrique du Sud, Chine, Israël, Russie 2004
  - Declaration on Access to Research Data from Public Funding
  - Demande à l'OCDE de formuler des principes et directives, L'OCDE se préoccupait de l'accès aux données de la recherche obtenues sur financement public
- OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007
  - Openness, flexibility, transparency, legal conformity, protection of intellectual property, formal responsibility, professionalism, interoperability, quality, security, efficiency, accountability, sustainability
- "factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings"

2013 : les principes du G8

Prise de position forte des Ministres de la Recherche du G8 en juin 2013

- i. To the greatest extent and with the fewest constraints possible **publicly funded scientific research data should be open**, while at the same time respecting concerns in relation to privacy, safety, security and commercial interests, whilst acknowledging the legitimate concerns of private partners.
- ii. Open scientific research data should be easily **discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards**.
- iii. To maximise the value that can be realised from data, the mechanisms for delivering open scientific research data should be efficient and cost effective, and consistent with the potential benefits.
- iv. To ensure successful adoption by scientific communities, open scientific research data principles will need to be underpinned by an appropriate policy environment, including **recognition of researchers** fulfilling these principles, and **appropriate digital infrastructure**.

2015-2016

- Octobre 2015, Berlin
  - Further progress on sharing and managing scientific data and information should be achieved, especially by **continuing engagement with community based activities such as the Research Data Alliance RDA**.
  - We encourage the GSO to continue their work on convergence and alignment of **inter-operable data management** that could accomplish an effective open-data science environment at the G7 level and beyond.
- Open Science statement – Entering into a new era for science (Mai 2016, Tsukuba)
  - Establish a working group on open science with the aims of **sharing open science policies**, exploring supportive **incentive structures**, and **identifying good practices** for promoting increasing access to the results of publicly funded research, including scientific data and publications, coordinating as appropriate with the Organisation for Economic Co-operation and Development (OECD) and Research Data Alliance (RDA), and other relevant groups; and
  - **Promote international coordination and collaboration** to develop the appropriate technology, infrastructure, including digital networks, and human resources for the effective utilization of open science for the benefit of all.

Les infrastructures de données

- Les données ne perdurent que dans des infrastructures identifiées
- Toutes les disciplines sont concernées : bio, environnement, SHS, astro, …
- Certains ESFRI sont des infrastructures de données : Elixir, Cessda, clarin,..

- Questions sur les données pour les infrastructures de recherche candidates à la Feuille de Route Nationale depuis la mise à jour 2016 (en cours de révision)
- Certaines IR sont des infrastructures de données : CDS, centres de ressources biologiques, biobanques, pôle de données pour le système terre, collections naturalistes, ECOSCOPE/biodiversité, PROGEDO/enquêtes,…

- Contexte européen : EOSC European Open Science Cloud
  - De l'agrégation de jeux de données aux services d'analyse et de traitement : outils logiciels, visualisation, moyens de calculs
  - H2020 work programme 2018-2020 Annexe 4 : développement des e-infrastructures

EOSC : le temps des réflexions

2016

European Commission

DIRECTORATE-GENERAL FOR RESEARCH AND INNOVATION (RTD)

New policy initiative: The establishment of an Open Science Policy Platform

Steering Group of the Policy Platform
(20 Members, co-chaired by the Commission, meets bi-annually)
Mandate: Develop the Open Science Policy Agenda; Promote uptake of agreed best policy practices

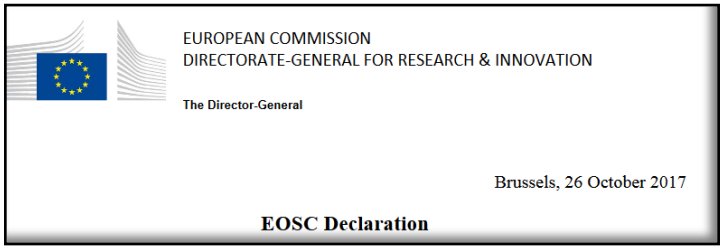
Working Group Rewards	Working Group Alt-Metrics	Working Group Research Integrity	Working Group Citizen Science	Working Group Open Education and skills	Working Group Changing Business Models	Working Group Fair Open Data	Working Group Open Science Cloud

FAIR : Findable, Accessible, Interoperable, Reusable

https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

EOSC: le temps des déclarations

- Sommet EOSC, juin 2017
  - data culture, data stewardship: practical and policy tools;
  - adoption and implementation of FAIR data principles;
  - research data infrastructures and services;
  - sustainable funding & governance;
  - High-performance computing, big data and super connectivity.


EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR RESEARCH & INNOVATION

The Director-General

Brussels, 26 October 2017

EOSC Declaration

- EOSC declaration, octobre 2017
  - data culture and FAIR data
    - Open access by default, skills, rewarding/incentives, trust, DMP, standards, legal aspects,…
  - Research data services and architecture
    - EOSC architecture, user needs, thematic specific needs, service deployment, EuroHPC
  - Governance and funding
    - Governance model, sustainability, coordinated co-funding, global aspects

EOSC : le temps de l'action

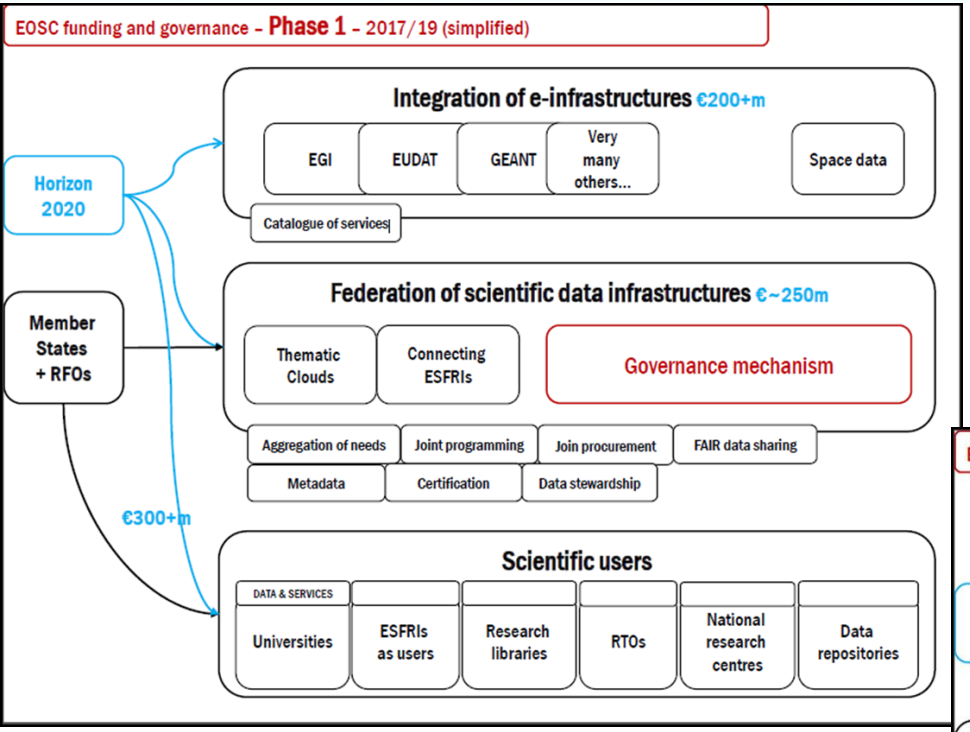
- Toutes les disciplines ne sont pas au même état d'avancement mais toutes sont concernées
- Tous les métiers ne sont pas au même niveau de sensibilisation
- Application du principe de subsidiarité :
  - gestion au plus près des besoins
  - fonctions de gouvernance, de financement à efficience maximale

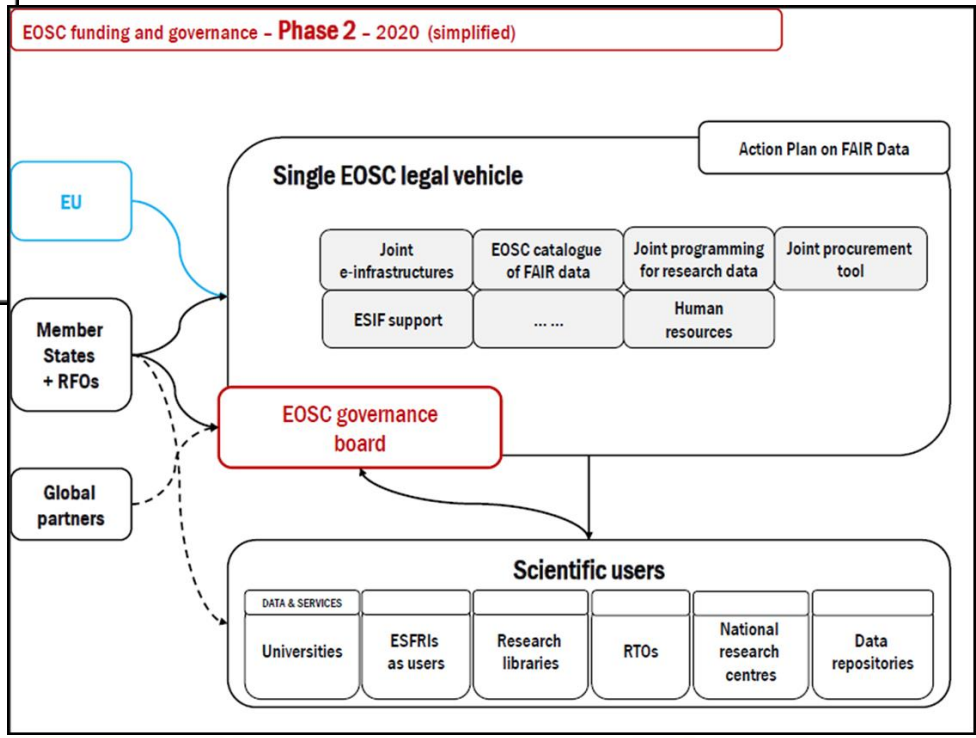
- Intégrer l'existant
  - ESFRI : data infrastructures
  - Les infrastructures génériques : GEANT, EGI, EUDAT, OpenAIRE,…
- The FAIR Guiding Principles for scientific data management – Findable, Accessible, Interoperable, Reusable http://www.nature.com/articles/sdata201618
  - Déjà réalisés dans certaines disciplines, en astronomie depuis très longtemps
  - Expert Group CE sur le comment faire

  http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3464
  - Initiative GO-FAIR
- L'émergence rapide de la Research Data Alliance (RDA)

  https://www.rd-alliance.org
- Recenser, harmoniser , labelliser les services déjà opérationnels
  - DMP, DOIs, ORCID, … pour faciliter l'accès, la publication et la citation des données
  - Services d'accompagnement : BBEES, DORANUM,…

EOSC : intégrer l'existant…





L’ infrastructure de données idéale…

- Un réservoir
  - Des réservoirs ?
  - Open, standards, confiance,
- Des services
  - Besoins utilisateurs (évolutifs)
  - Un pilotage opérationnel : qualité, confiance
  - FAIR data, open source
  - Traitement, analyse, visualisation, extraction, …
  - compétences
- Une gouvernance
  - Durable, co-construite, agile
- Un financement
  - Pérenne, efficient, partagé

The evolving landscape of Federated Research Data Infrastructures : a Knowledge Exchange report

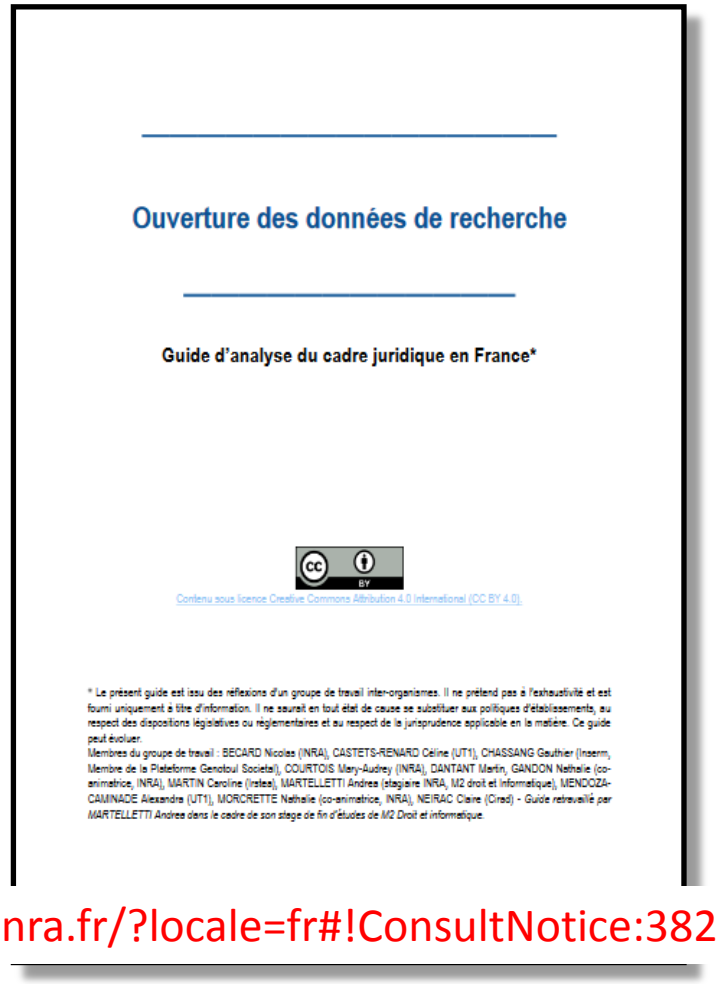
- KE ? A collaboration between DFG, CSC, Jisc, CNRS, SURF, Deff (Denmark's Electronic Research Library)
- Open scholarship, open access
- **" Essentially, a federated infrastructure is one where a range of distributed services are coordinated by an overarching level"**
  - *Associer les données ET les services (qualité/confiance, métadonnées, logiciels, visualisation)*
- **"Two broad sets of factors drive the emergence and development of federated infrastructures: push factors, which might be also be characterised as top-down; and demand from users, reflecting a bottom-up approaches"**
  - *Push: défis sociétaux, politique nationale, agences publiques (open data), financeurs, RPO's*
  - *Bottom-up : demandes culturelles fortes (astro, HEP, envi) ou moins fortes (SHS)*
- **"The involvement of users is also a crucial imperative, and infrastructures are careful to nurture their relationships with numerous partners within the academic sector and beyond"**
- **A major challenge for the development of federated infrastructures is the complexity and fragmented nature of the research data environments in which they evolve**
  - *Hétérogénéité des cadres légaux et réglementaires, des systèmes administratifs, des régimes de financements, … et ce au sein de chaque discipline à l'échelle de l'Europe*
  - *Lent processus : changement culturel, stabiliser le financement, assurer compatibilité technique, …*
- **The emergence of EOSC is generally welcomed, particularly since it is seen as reflecting the same rationale as national infrastructures, albeit at a pan-European scale – with the beneficial scaling up that this could imply**
  - *Besoins des utilisateurs au centre du dspositif, gouvernance solide, claire répartition des taches*
  - *Se développer sur les services déjà existants*

Contexte juridique
la science est mal servie par le législateur

- Un contexte en évolution…
- Des législations nationale et européenne pas toujours en harmonie…

- Loi CADA, modifiée par la loi Valter (2015)
- Transposition de la directive PSI
- Loi Pour une république numérique (2016)
- A venir, révision de la directive européenne sur les droits d'auteurs et droits voisins 2001/EC/29
- Mai 2018, Règlement général relatif à la protection des données

Et si on inscrivait les principes de l'EOSC dans la loi ?



Ouverture des données de recherche

Guide d'analyse du cadre juridique en France*

* Le présent guide est issu des réflexions d'un groupe de travail inter-organismes. Il ne prétend pas à l'exhaustivité et est fourni uniquement à titre d'information. Il ne saurait en tout état de cause se substituer aux politiques d'établissements, au respect des dispositions législatives ou réglementaires et au respect de la jurisprudence applicable en la matière. Ce guide peut évoluer.

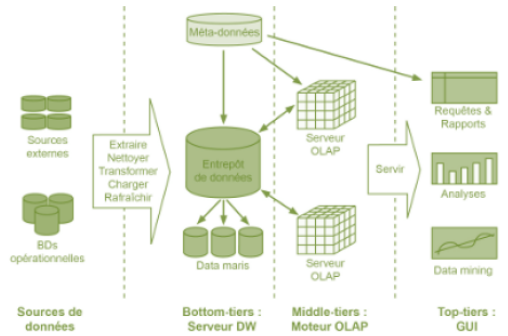
Membres du groupe de travail : BECARD Nicolas (INRA), CASTETS-RENARD Céline (UT1), CHASSANG Gauthier (Inserm, Membre de la Plateforme Genotoul Societal), COURTOIS Mary-Audrey (INRA), DANTANT Martin, GANDON Nathalie (co-animatrice, INRA), MARTIN Caroline (Irstea), MARTELLETTI Andrea (stagiaire INRA, M2 droit et informatique), MENDOZA-CAMINADE Alexandra (UT1), MORCRETTE Nathalie (co-animatrice, INRA), NEIRAC Claire (Cirad) - *Guide retravaillé par MARTELLETTI Andrea dans le cadre de son stage de fin d'études de M2 Droit et informatique.*

http://prodinra.inra.fr/?locale=fr#!ConsultNotice:382263

Propriété des données

- La donnée de recherche, une information de « libre parcours »
- Obligation de diffuser les données… mais de nombreuses exceptions.

Données de la recherche, essai de définition



Ce qu'il faut retenir :

- Pas de distinction entre données brutes, élaborées ou métadonnées d'un point de vue juridique.
- Pas de droit de propriété dans la plupart des cas sur la donnée (données machine, etc.). Elle est considérée comme une information « de libre parcours ». A ce titre, l'établissement du producteur de la donnée peut restreindre ou non sa diffusion.
- Mais il existe deux exceptions où une « propriété » peut s'exercer.

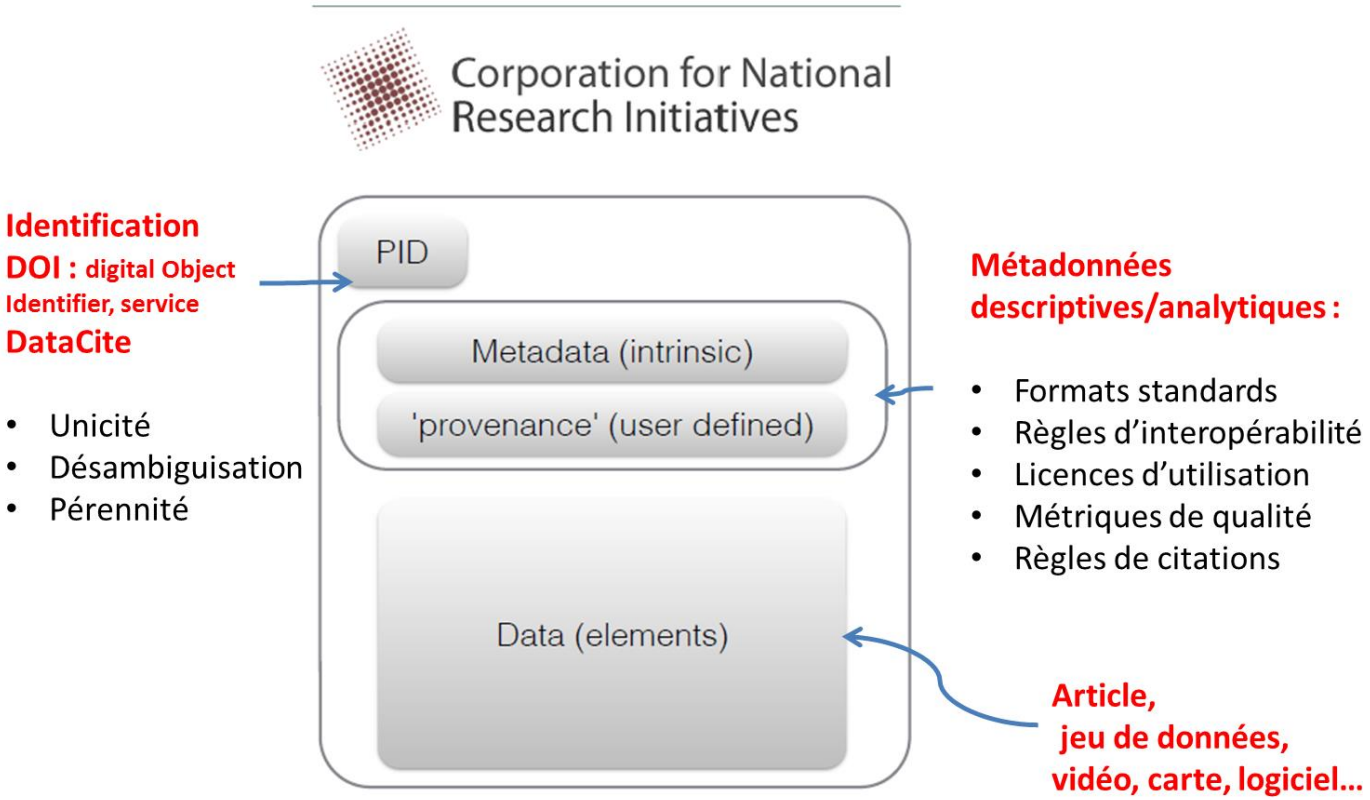
Des exceptions en fonction de la nature des données

Des interdictions totales de diffusion :

- Les documents réalisés en exécution d'un contrat de prestation de services exécuté pour le compte d'une ou de plusieurs personnes déterminées (non publiques)
- Les données relevant du secret défense
- Les données relatives aux secrets professionnels : secret des procédés, secret des informations économiques et financières, secret des stratégies commerciales ou industrielles
- Données portant atteinte à la sécurité du SI de l'administration (NOUVEAU).

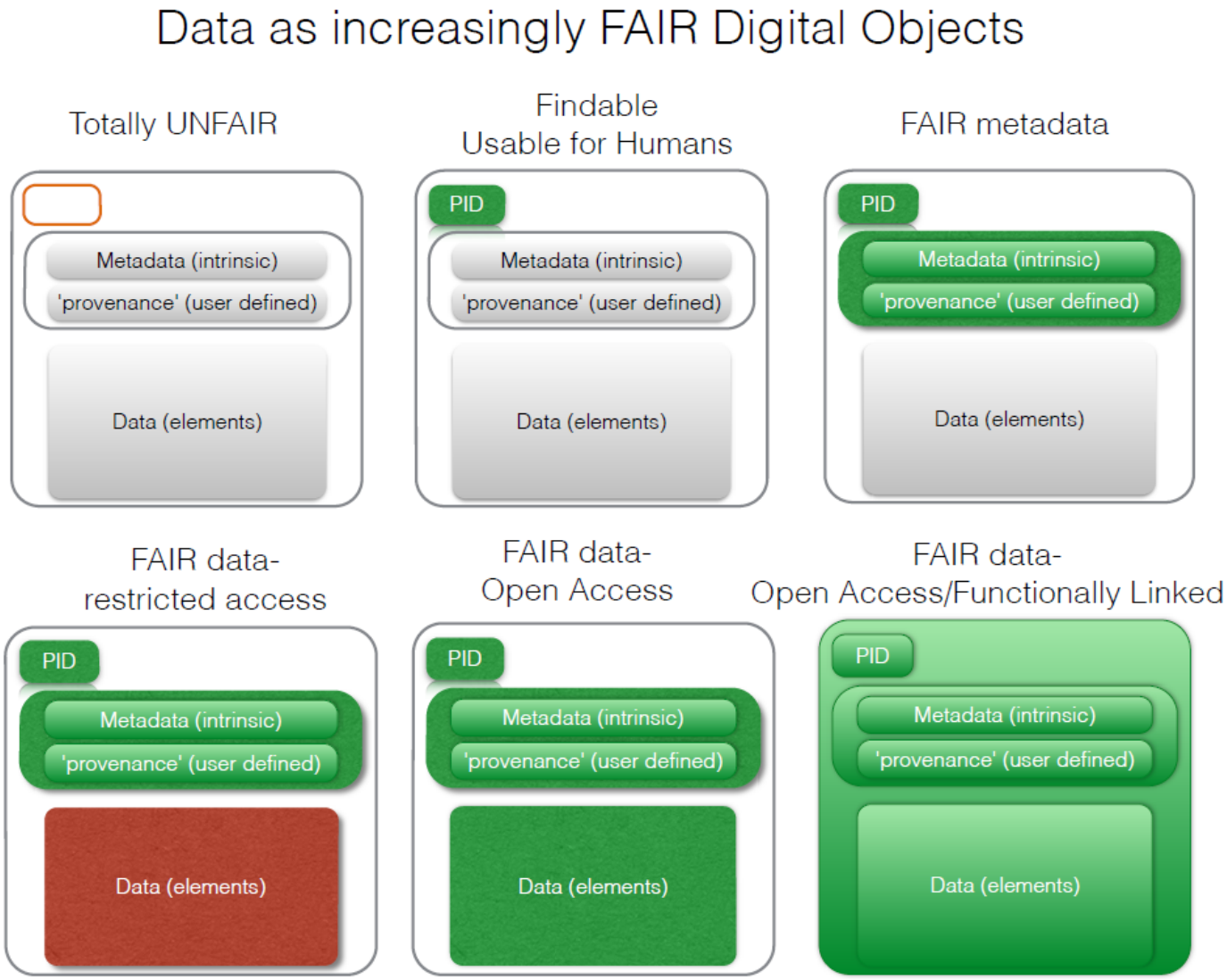
Nathalie Gandon, Nathalie Morcrette,Inra

Findable, Accessible, Interoperable, Reusable

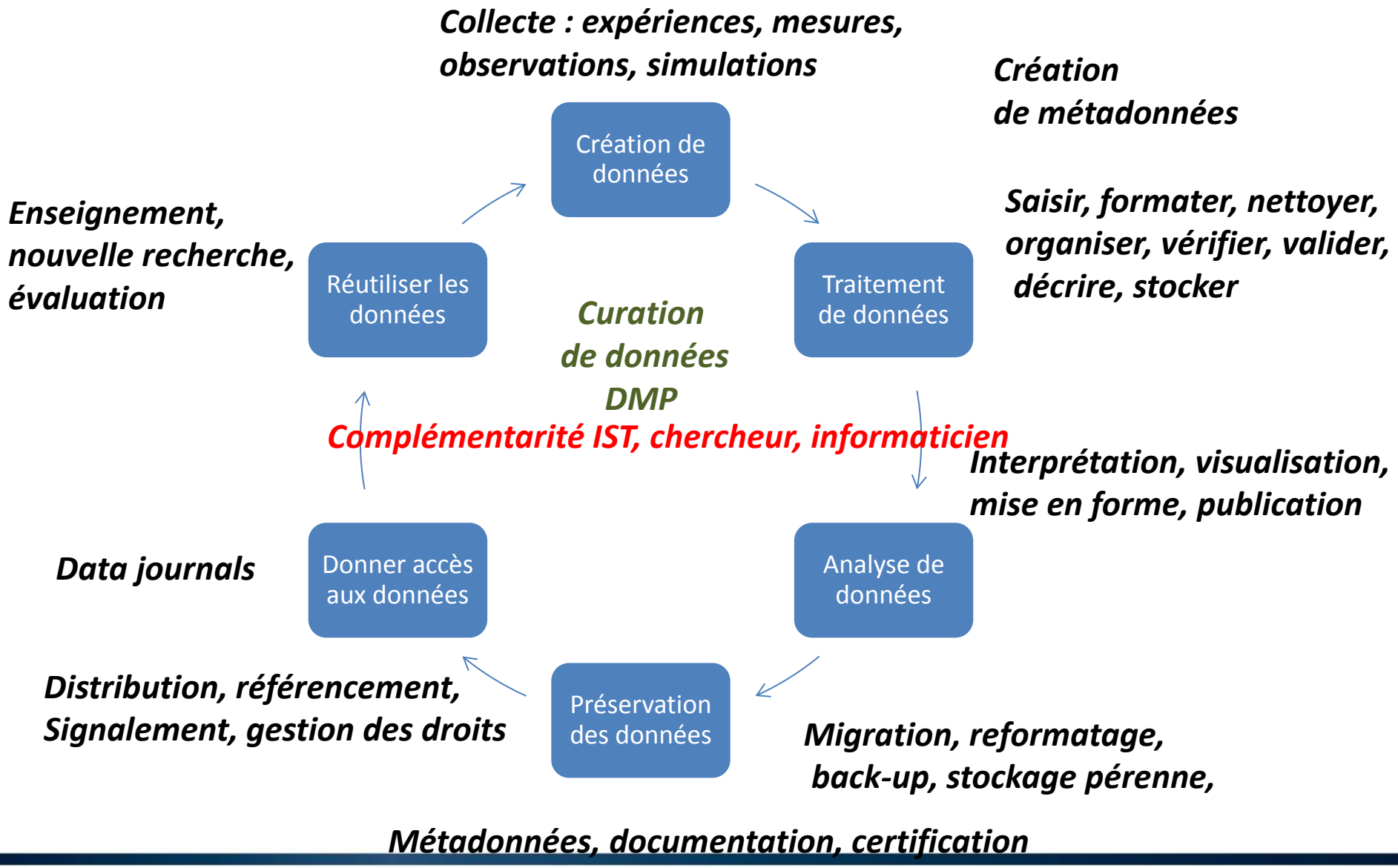


A simplified diagram of a Digital (data) Object irrespective of technological choices and naming

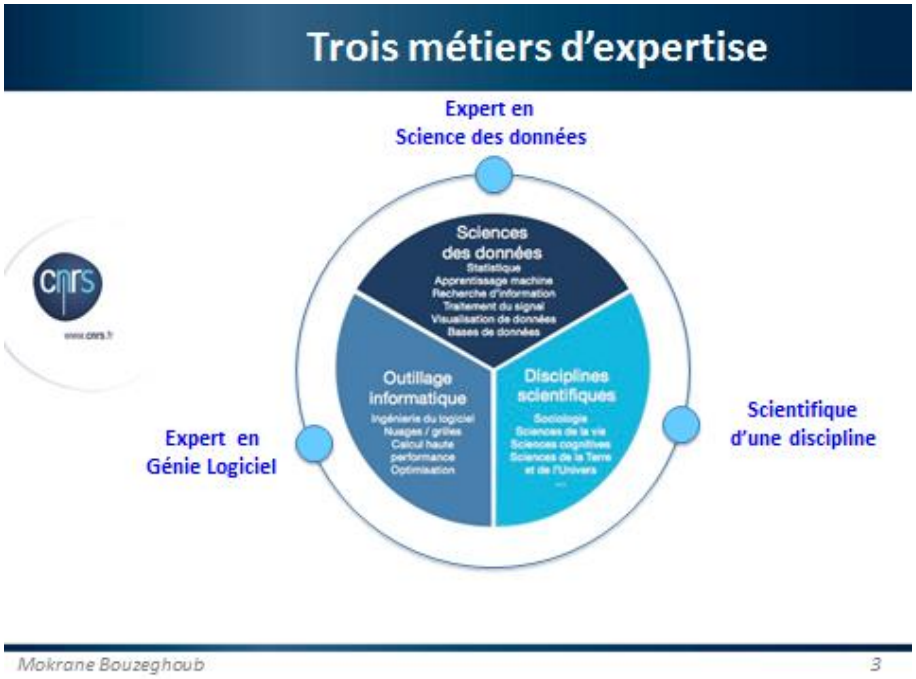
FAIR principles transition



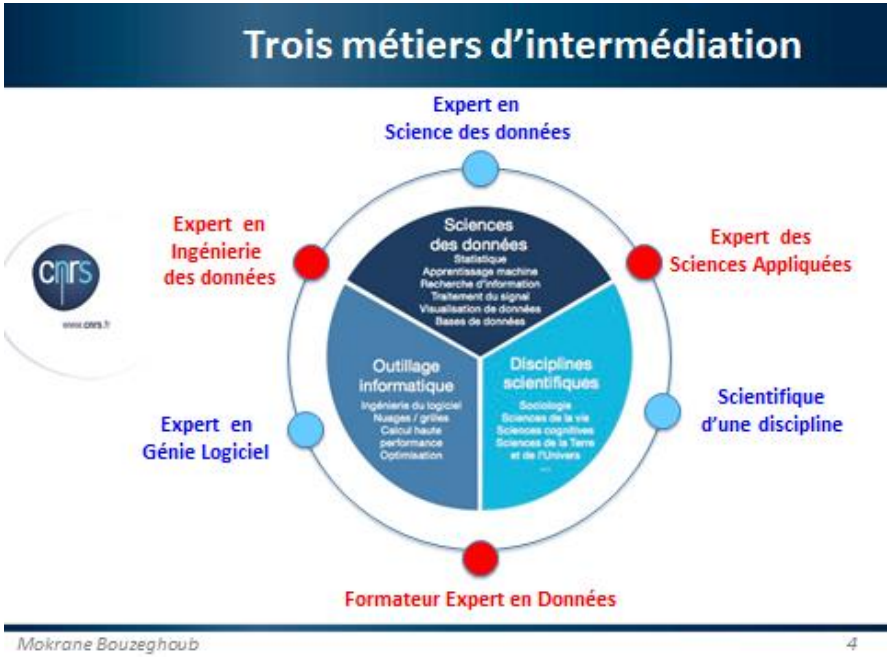
Favoriser la curation des données



Data science, data management



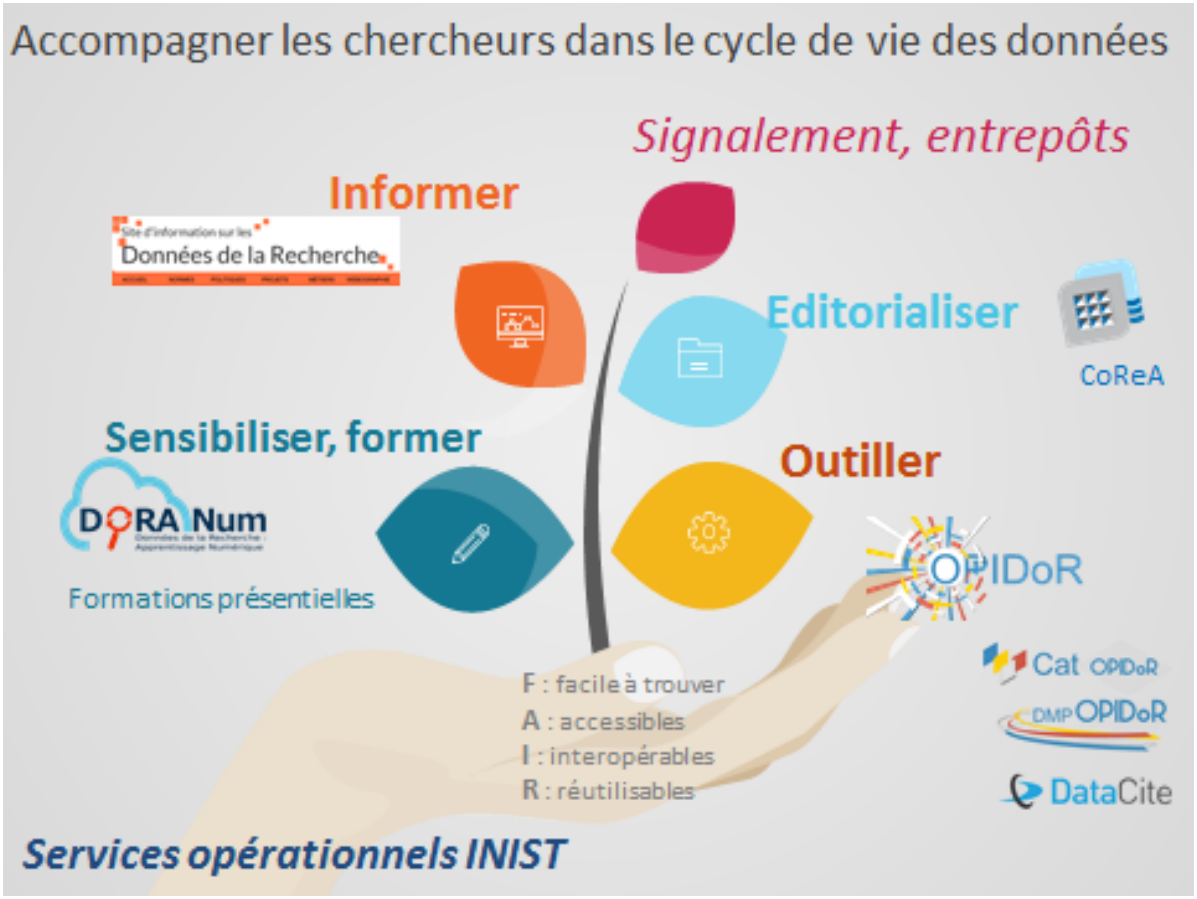
- Des compétences complémentaires pour l'exploitation ET la gestion des données
- Le temps du chercheur n'est pas (encore) le temps de la donnée

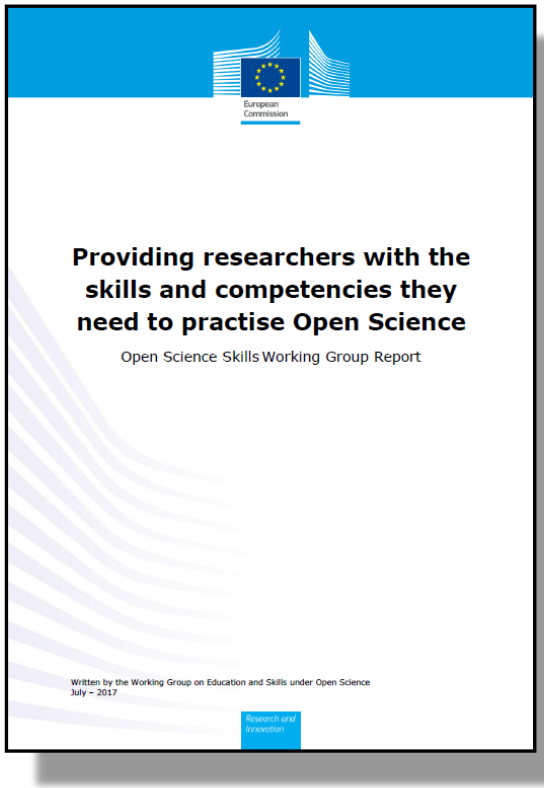


- Les métiers d'intermédiation sont essentiels pour assurer aux données une capacité de réutilisation dans un autre contexte. (interdisciplinarité)

d'après M. Bouzeghoub, présentation oct 2017

Sensibilisation, développement des compétences





https://ec.europa.eu/research/openscience/pdf/os_skills_wgreport_final.pdf

Stratégie(s) institutionnelle(s)

- S’inscrire dans un contexte international et européen
- Respecter les pratiques disciplinaires
- Porter les meta-messages :FAIR principles, Open access by default, data culture
- Favoriser le développement des compétences
- Opérer les infrastructures de données
- Pérenniser les efforts de gestion des données, au-delà des projets qui les ont générées

Les invariants : « take home messages »

- Les données (FAIR data) sont un élément essentiel de la politique européenne d'Open Science
- Les données sont des objets précieux :
  - Valorisation du chercheur et de son institution
  - Administration de la preuve
  - Réutilisation, interdisciplinarité

- Les données ne sont partageables que si bien documentées (curation)
- Data scientist et data curator, des métiers d'avenir !
- La confiance (qualité) est le gage d'une (possible) réutilisation
- Les barrières sociales (collectives et individuelles) prennent le pas sur les contraintes techniques
- Les scientifiques sont les garants des bonnes pratiques de gestion des données, et de leur réutilisation. A ce titre, ils doivent participer à la gouvernance des infrastructures globales type EOSC

Merci de votre attention

francis.andre@cnrs-dir.fr

A voir….

Moving Towards Plenary 11: Berlin!

DU à l'UPMC : Data Stratégie http://www.data-strategie.upmc.fr/