DATALABO 2017

GÉRER LES DONNÉES DE LA RECHERCHE :
PARTAGE DE BONNES PRATIQUES

MARDI 28 MARS 2017
AUDITORIUM IMAG - GRENOBLE
9H - 17H

Données de la recherche: vers de bonnes pratiques de gestion dans l'environnement national et international

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

Le contexte du partage des données: Evolution des pratiques de science

Science du 21ème siècle : plus…

- Numérique
- Collaborative
- Interdisciplinaire
- Réactive
- Citoyenne
- Partagée

Open Research

Science 2.0

eScience

OPEN SCIENCE

Tendances : accroissement de la production scientifique, du nombre de chercheurs, nouvelle façon de faire de la science, guidée par les données massives, importance des défis sociétaux

Culture du partage de la donnée

- Mes données sont à moi et … tes données sont à moi également !
- Des initiatives : ScienceEurope, Knowledge Exchange, LERU, LIBER, DCC, Nactem, RDA,…
- Des pratiques de communautés : astrophysique, génomique,…
- Des politiques d'organismes : Inra, Irstea,…
- Des politiques d'infrastructures
- Des politiques de financeurs : ANR, CE:H2020

- De nombreux niveaux d'intervention, besoins de pragmatisme et de dialogues entre les acteurs >>>> Research Data Alliance

RDA lancé en 2012

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE

Vision:

Researchers openly **sharing data** across technologies, disciplines, and countries

Mission:

Building the social and technical **bridges**

Guiding Principles

- Openness
- Consensus
- Balance
- Harmonization
- Community-driven
- Non-profit

Research Data Alliance created to Accelerate Development of Research Data Sharing Infrastructure Worldwide

- RDA community focuses on building **social, organizational and technical infrastructure** to
  - reduce barriers to data sharing and exchange
  - accelerate the development of coordinated global data infrastructure

CREATE → ADOPT → USE

RDA Working Group Infrastructure Deliverables are:

- **Focused pieces of adopted code, policy, infrastructure, standards, or best practices** that enable data to be shared and exchanged
- **"Harvestable" efforts** for which 12-18 months of work can eliminate a roadblock for a substantial community
- **Efforts that have substantive applicability** to "chunks" of the data community, but may not apply to everyone
- **Efforts for which working scientists and researchers can start today** while more long-term or far-reaching solutions are appropriately discussed in other venues

RDA : culture du partage de la donnée

Research Data Sharing without barriers

RESEARCH DATA ALLIANCE

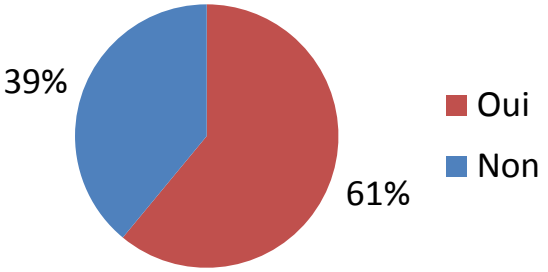
https://rd-alliance.org/sites/default/files/attachment/Booklet_OutputsSeptember2015_web.pdf

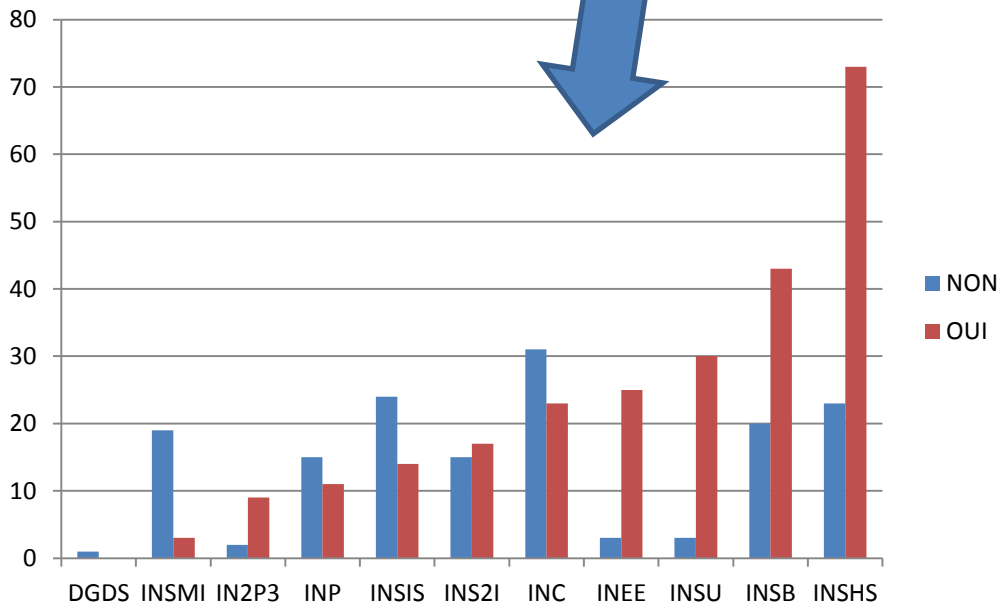
- **Australian Commonwealth Government** through the **Australian National Data Service** supported by the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative ;
- **European Commission** through the RDA Europe project funded under the **7th Framework Program** ;
- **United States of America** through the RDA/US activity funded by the **National Science Foundation** and other U.S. agencies.

Enquête IST auprès des directeurs d'unité

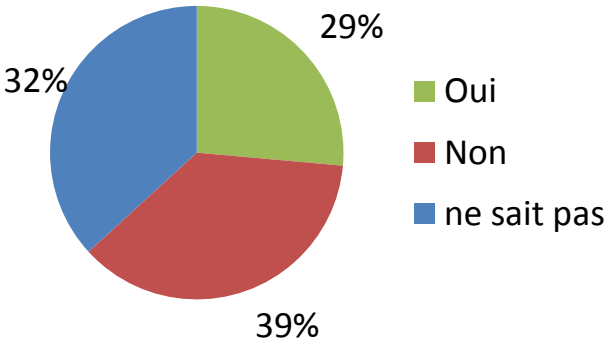
Données de la recherche

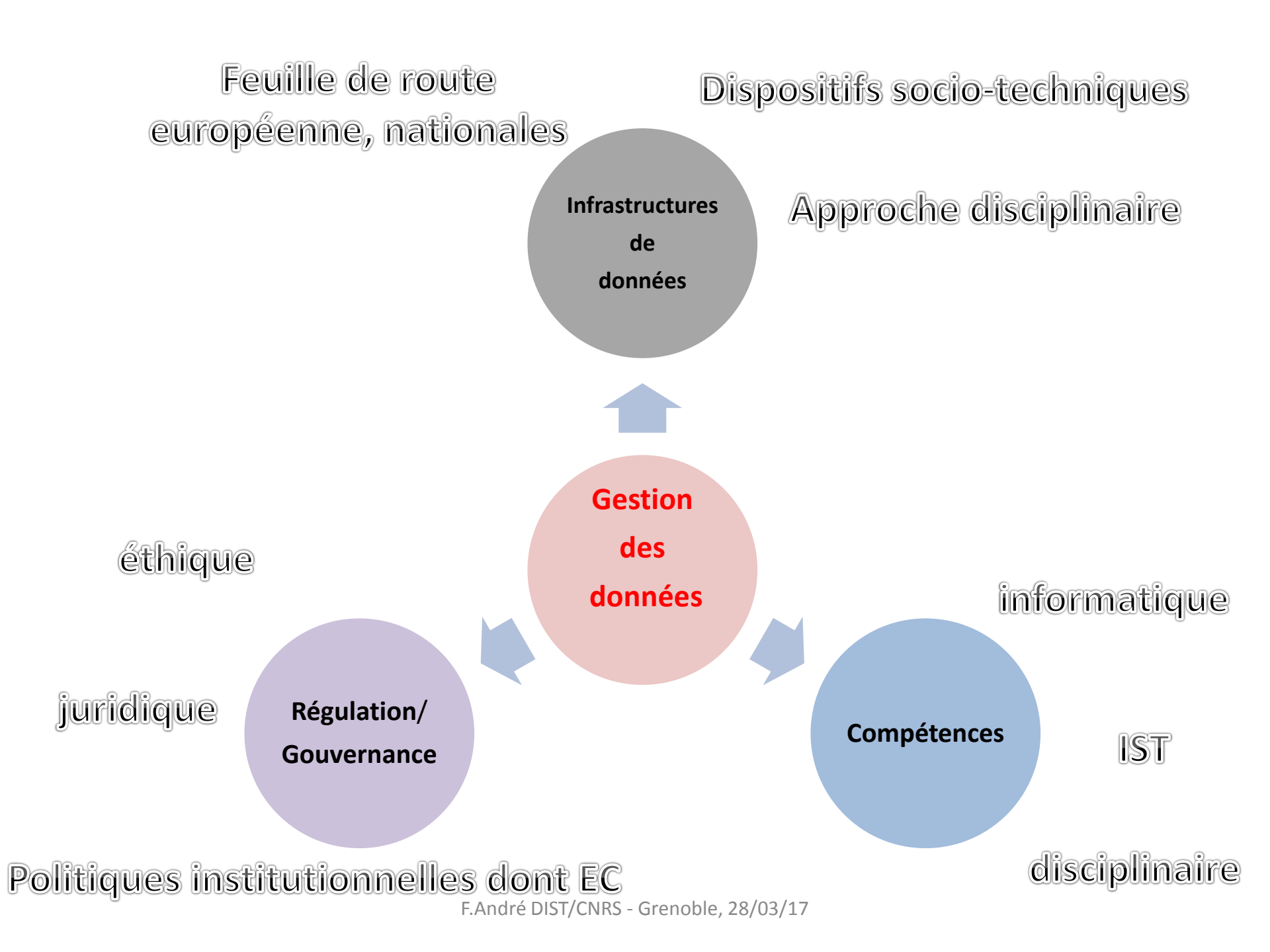
Les recherches conduites dans votre laboratoire produisent-elles des données de la recherche nécessitant des pratiques de gestion (description, archivage, diffusion…) ?

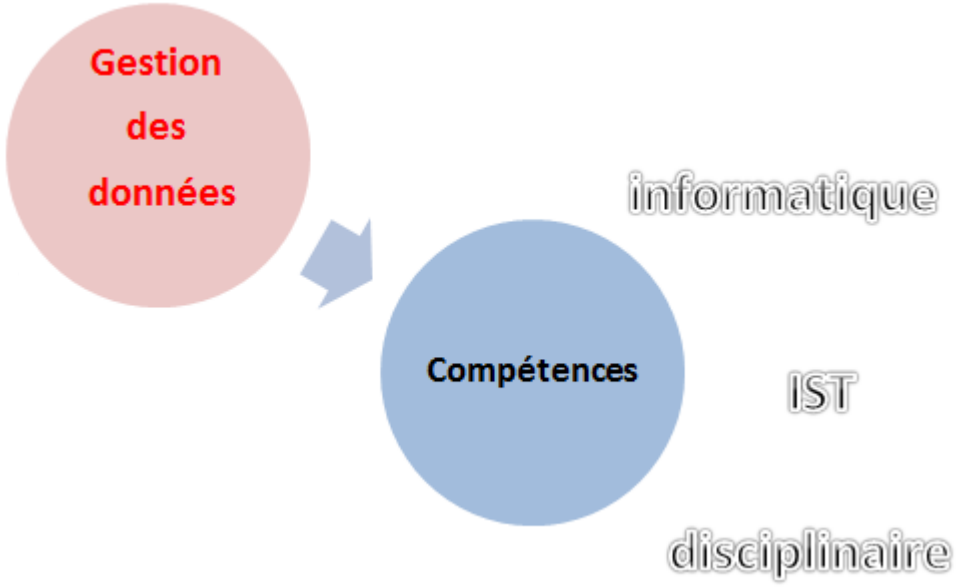




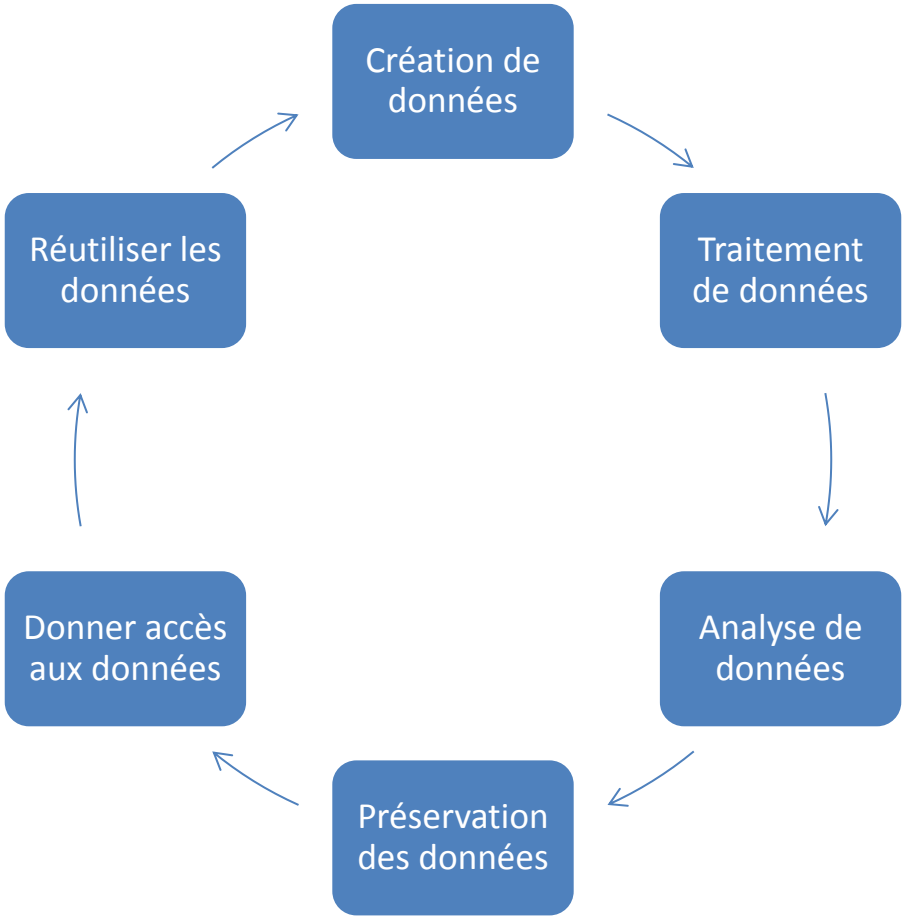
Pensez-vous que vos données de recherche soient libres de droits ?



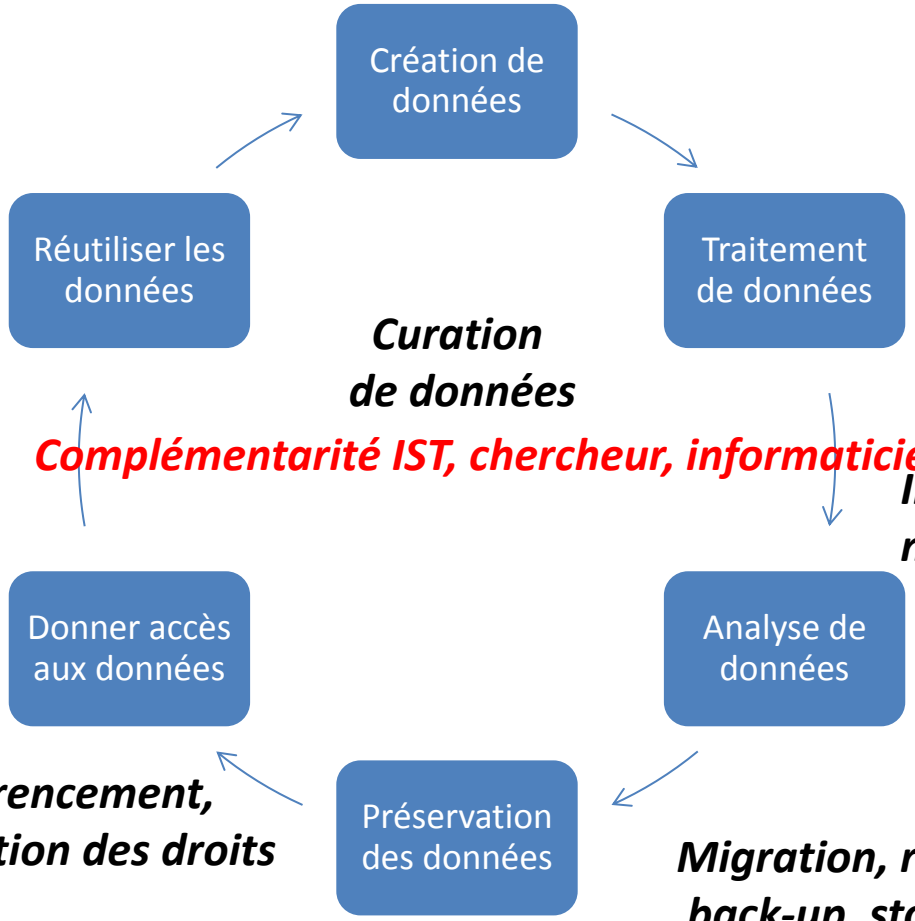

Feuille de route européenne, nationales
Dispositifs socio-techniques
Infrastructures de données
Approche disciplinaire
Gestion des données
éthique
informatique
juridique
Régulation/ Gouvernance
Compétences
IST
Politiques institutionnelles dont EC
disciplinaire


Gestion
des
données
Compétences
informatique
IST
disciplinaire

Cycle des données



Cycle des données



Collecte : expériences, mesures, observations, simulations

Création de métadonnées

Saisir, formater, nettoyer, organiser, vérifier, valider, décrire, stocker

Interprétation, visualisation, mise en forme, publication

Migration, reformatage, back-up, stockage pérenne,

Métadonnées, documentation, certification

Distribution, référencement, Signalement, gestion des droits

Data journals

Enseignement, nouvelle recherche, évaluation

Curation de données

Complémentarité IST, chercheur, informaticien

On recherche…

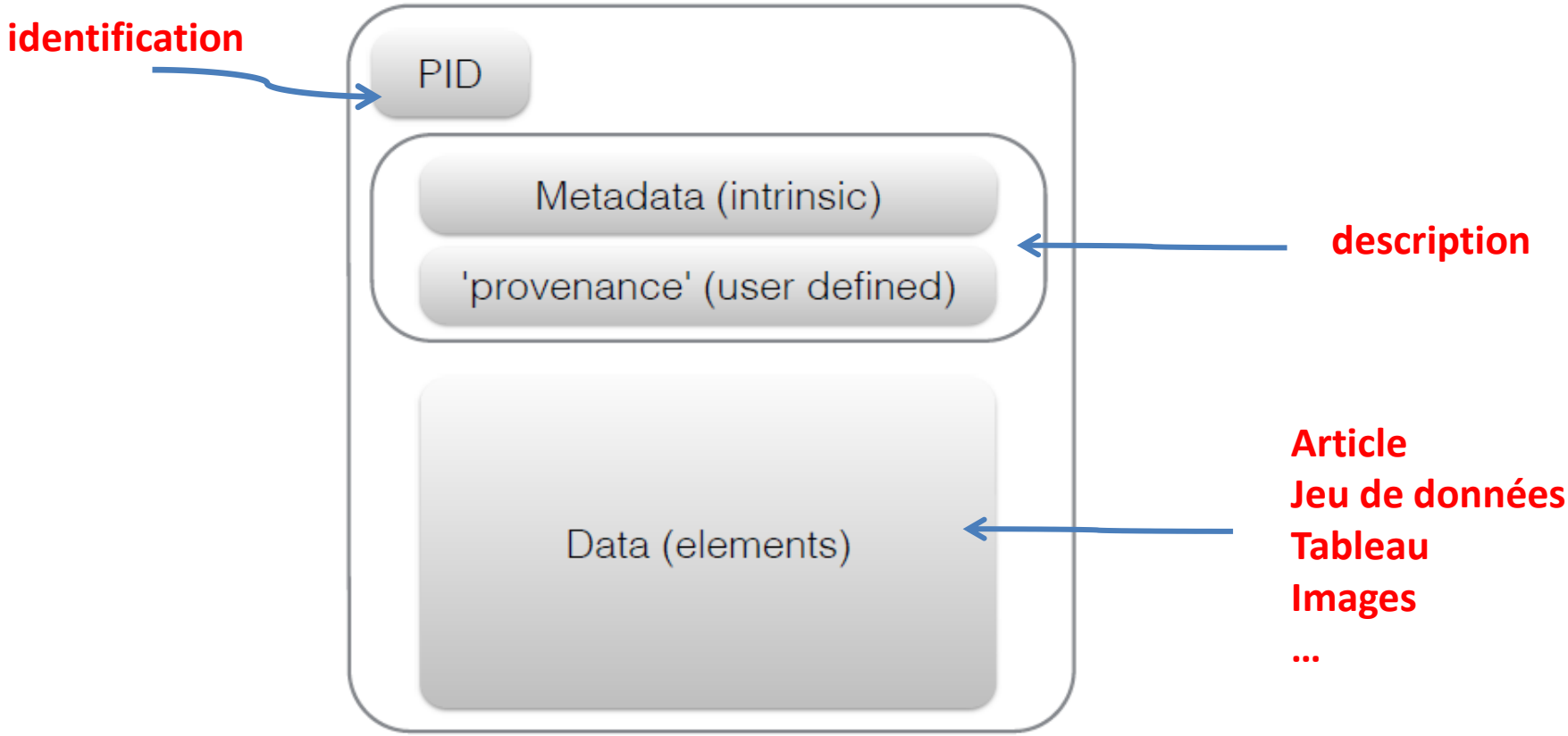
Data Experts: this term refers to a distinct and largely novel class of research professional. They are not traditional core computer or data scientist, but embedded data specialists that are able to support domain specific researcher throughout the entire knowledge discovery cycle. They typically do not end up with high impact factors in traditional systems but should become indispensable core partners in any modern data driven research team with a solid perspective.

HLEG EOSC report oct. 2016

Data scientist, data librarian, data analyst, data curator, ….

http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

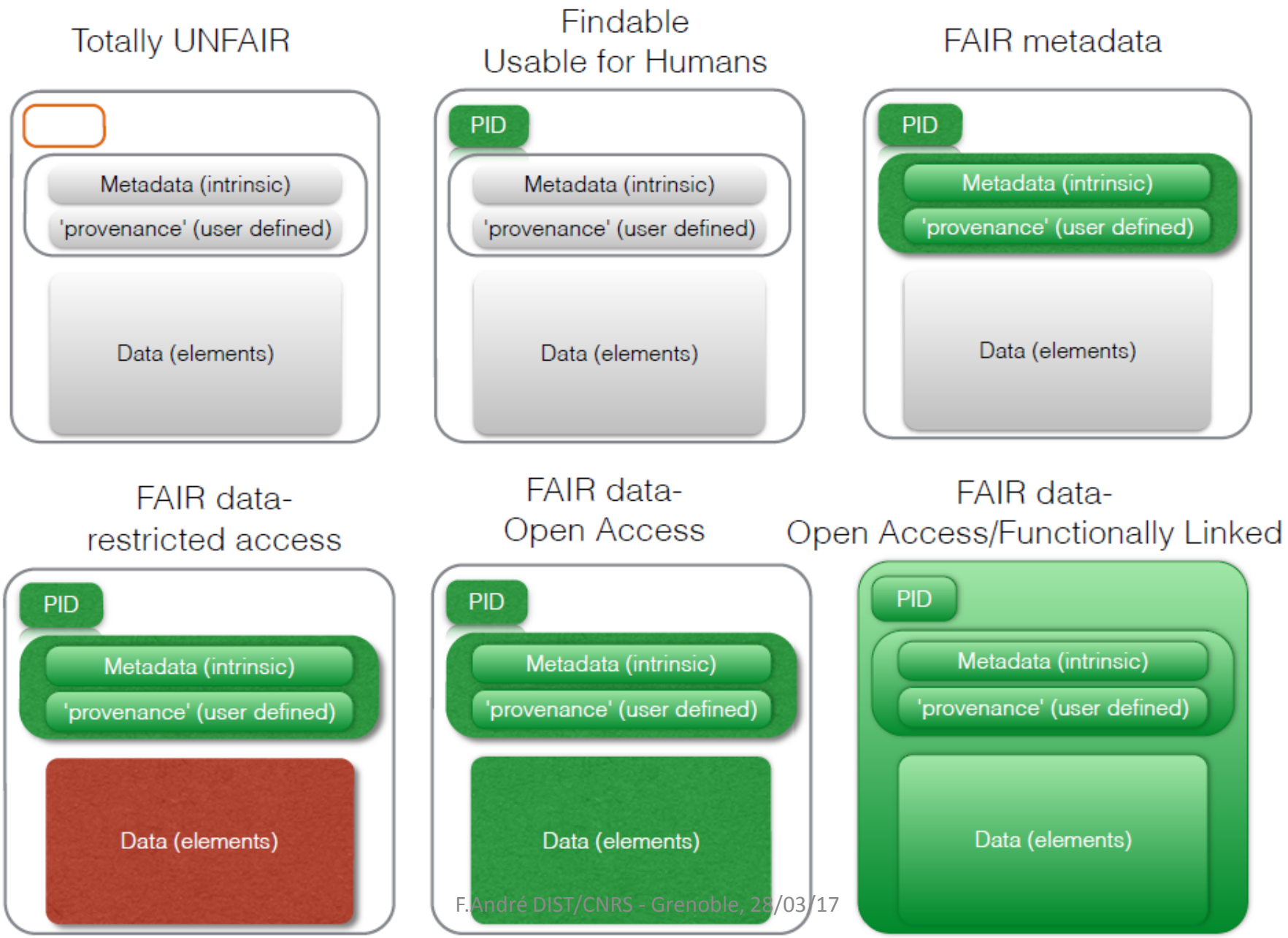
Corporation for National Research Initiatives



Soyez **FAIR** ! : **F**indable, **A**ccessible, **I**nteroperable, **R**eusable

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

Data as increasingly FAIR Digital Objects

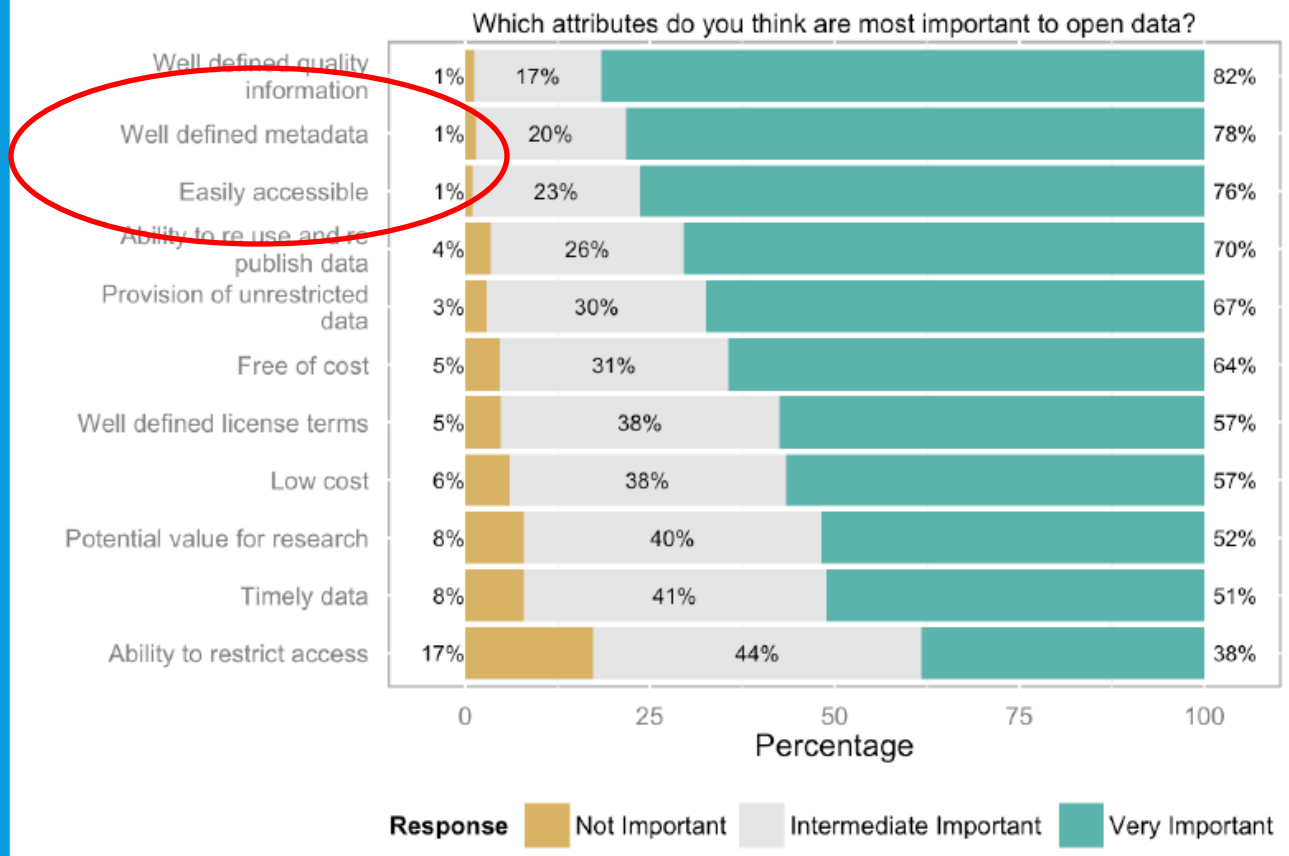


Où se niche la qualité des métadonnées ?

- Richesse des métadonnées lisibles en machine
- Identifiants pérennes
- Variété et disponibilité de formats
- Règles d'interopérabilité documentées
- Licences publiées
- Métriques de qualité affichées
- Mises à jour des métadonnées
- Règles de citations

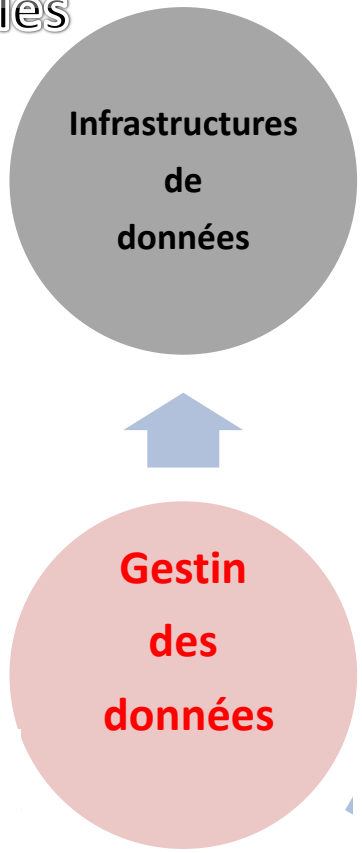
Données ouvertes : caractéristiques attendues

Qualité des données et des métadonnées



Feuille de route européenne, nationales

Dispositifs socio-techniques

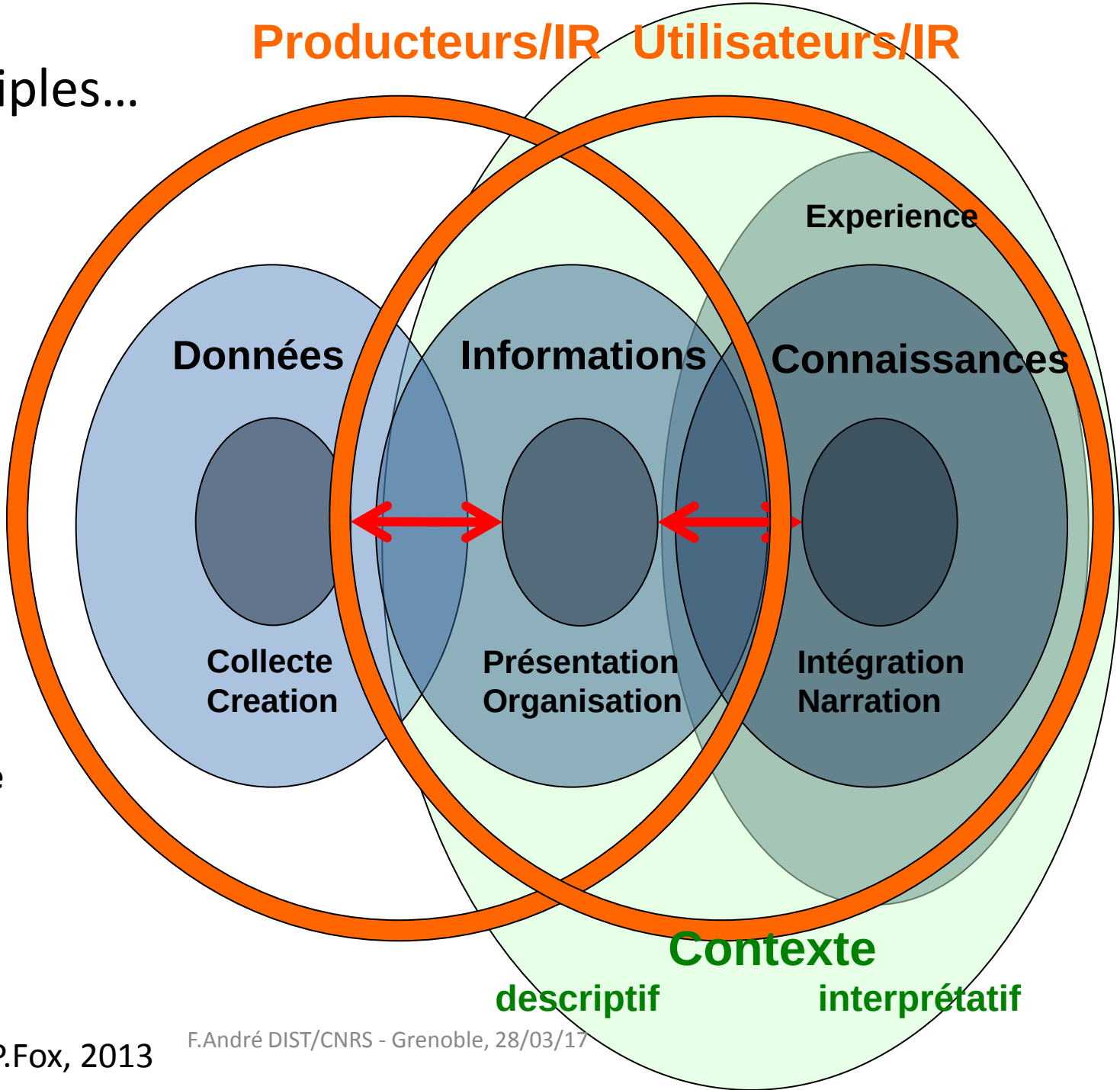


Approche disciplinaire

Acteurs multiples…

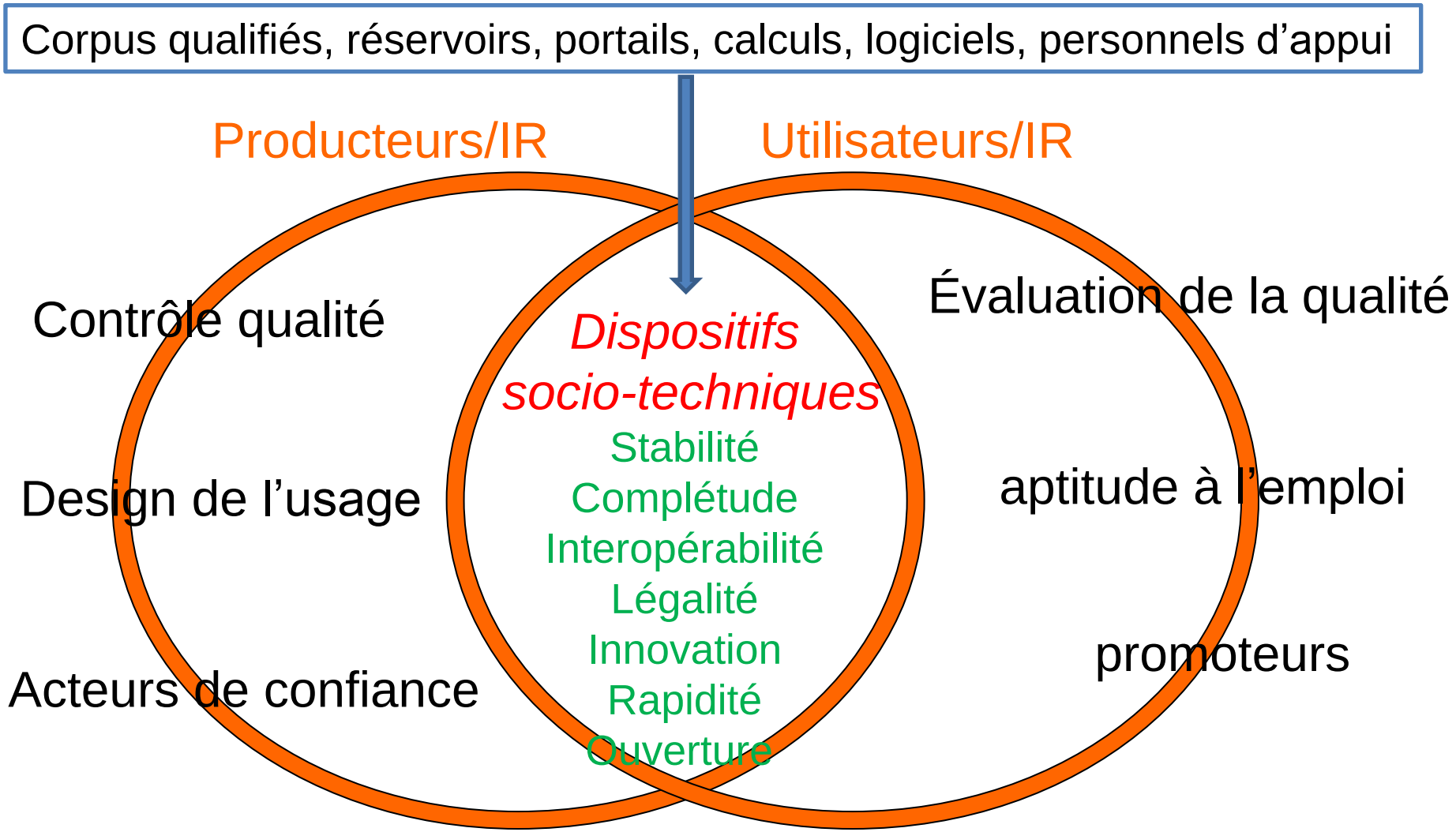
Du processus de recherche…

à la construction de l'infrastructure



Adapté d'après P.Fox, 2013

IR, curation et usages



Trust, Trust, Trust !

Banal aujourd'hui :

- Certifications des publications et des données
  - Peer reviewed journals and data journals
  - Les archives institutionnelles : preprints/postprints
  - Open peer reviewing
  - Le couplage publication/données

Moins habituel :

- Certifications des réservoirs de données : workflow, format, procédures qualité,…

Open Science à l'échelle européenne

- Quitter une situation de blocage
  - changer les processus d'évaluation
  - Modifier les règles de PI
  - Faciliter le TDM
  - Changer les modèles économiques de la diffusion de la science
- Promouvoir des politiques de science ouverte
  - Adopter (et adapter !) des principes d'accès libre
  - Stimuler les pratiques de recherche et d'innovation basées sur les données
- Développer des infrastructures de recherche
  - Basées sur des principes de partage
  - Mutualisées
- Impliquer les acteurs de la recherche
  - Chercheurs, personnels de soutien, société
  - Former, former, former…

http://www.eu2016.nl/documenten/rapporten/2016/04/04/amsterdam-call-for-action-on-open-science

Vision CE : les bénéfices du partage des données "Great opportunities for the society"

- Better value for money
  - By **strengthening the productivity of the European science** and research system through the uptake of results by businesses, in particular SMEs that may not have the resources to pay for access to research results
- More transparency, openness and collaboration
  - leading to a **higher degree of responsiveness** of the research community to societal challenges
- A sound science and society relationship
  - More openness may also lead to more **trustworthy science** from the point of view of the citizen and civil society organisations (NGOs)
- Big and open data are estimated to add 1.9% of EU-28 GDP by 2020

D'après J.F. Dechamp, CE, Directorate-General for Research & Innovation

Vision CE : les bénéfices du partage des données "Great opportunities for researchers"

- Wider dissemination and sharing of the results
- Involvement in **more interdisciplinary research**
- **More visibility and credit** for those collecting and sharing underlying research data
- Involvement in **international networks** full of potential
- **New career paths** e.g. data scientists, start-ups, science diplomacy

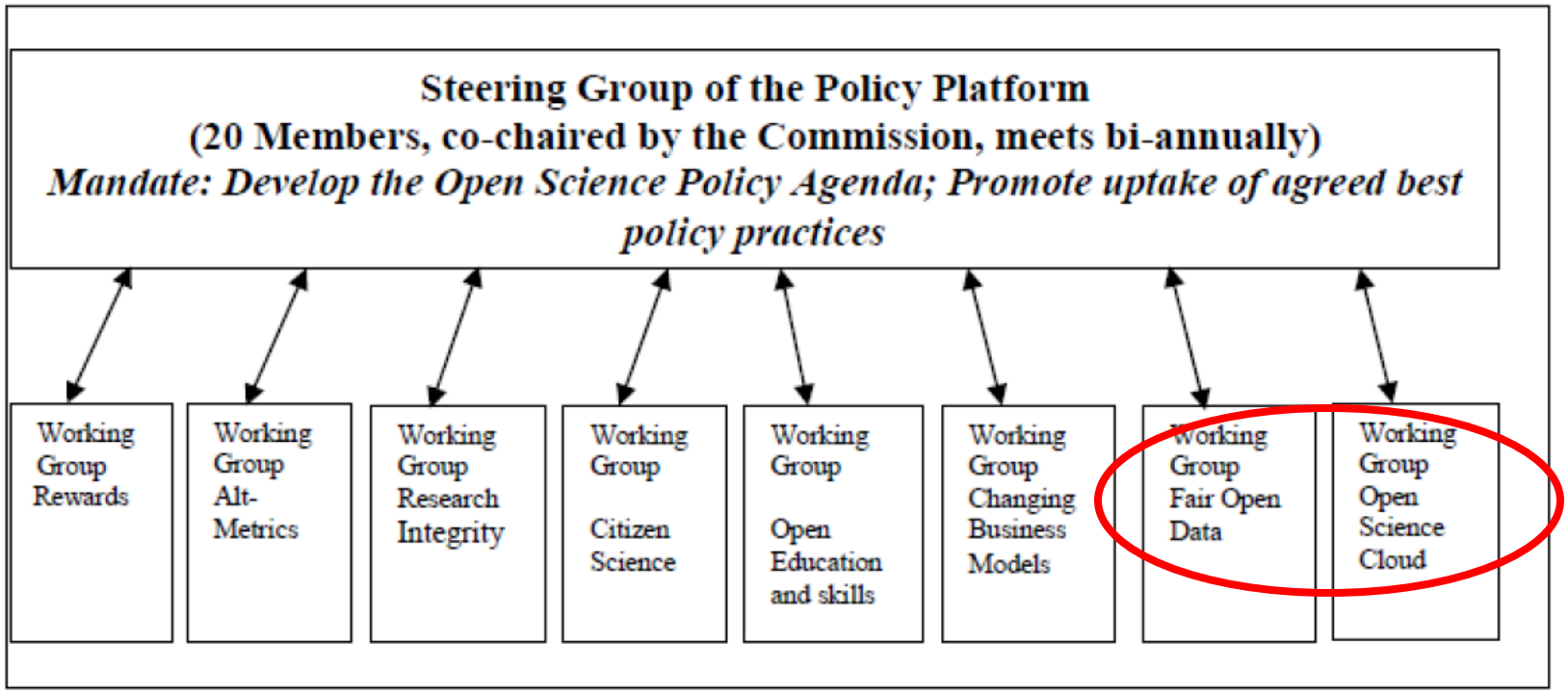
D'après J.F. Dechamp, CE, Directorate-General for Research & Innovation

2016

European Commission

DIRECTORATE-GENERAL FOR RESEARCH AND INNOVATION (RTD)

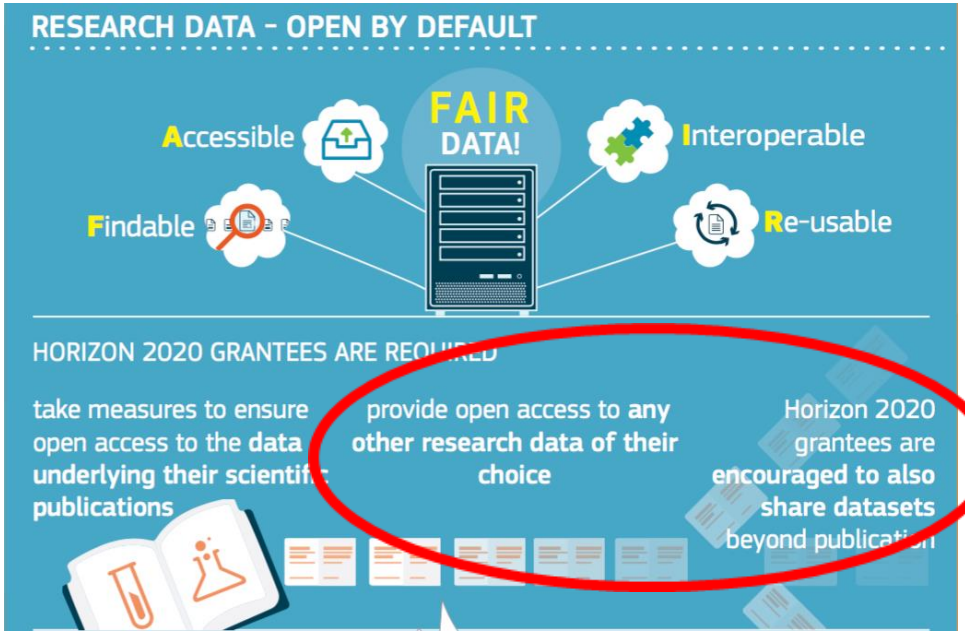
New policy initiative: The establishment of an Open Science Policy Platform

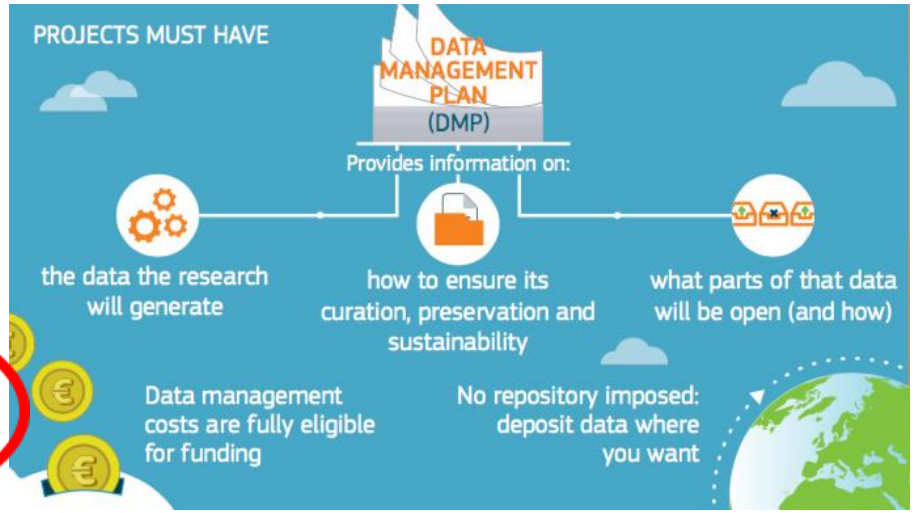


FAIR : Findable, Accessible, Interoperable, Reusable

Research data : Open by default

D'après Barend Mons, EOSC

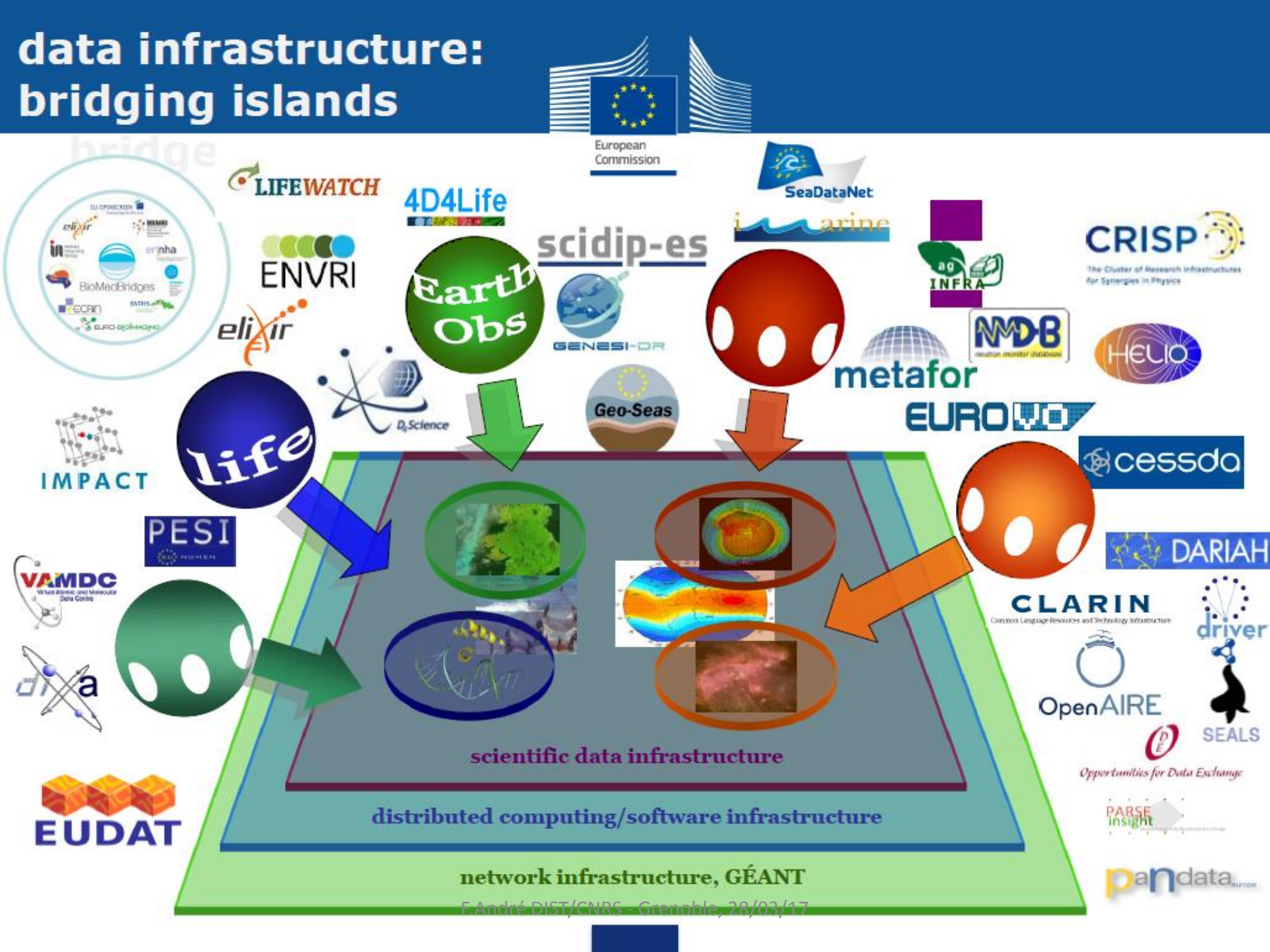


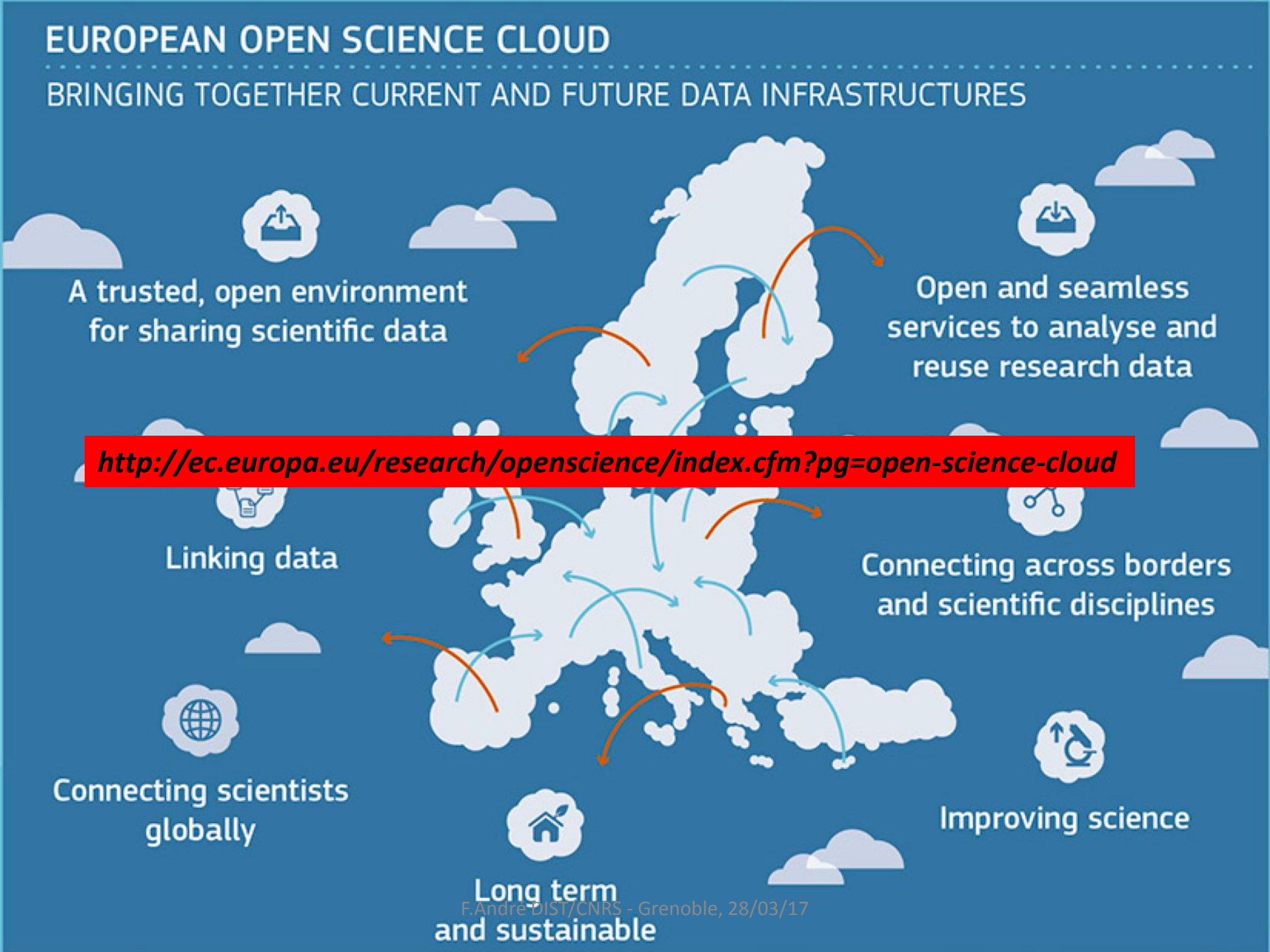


5%

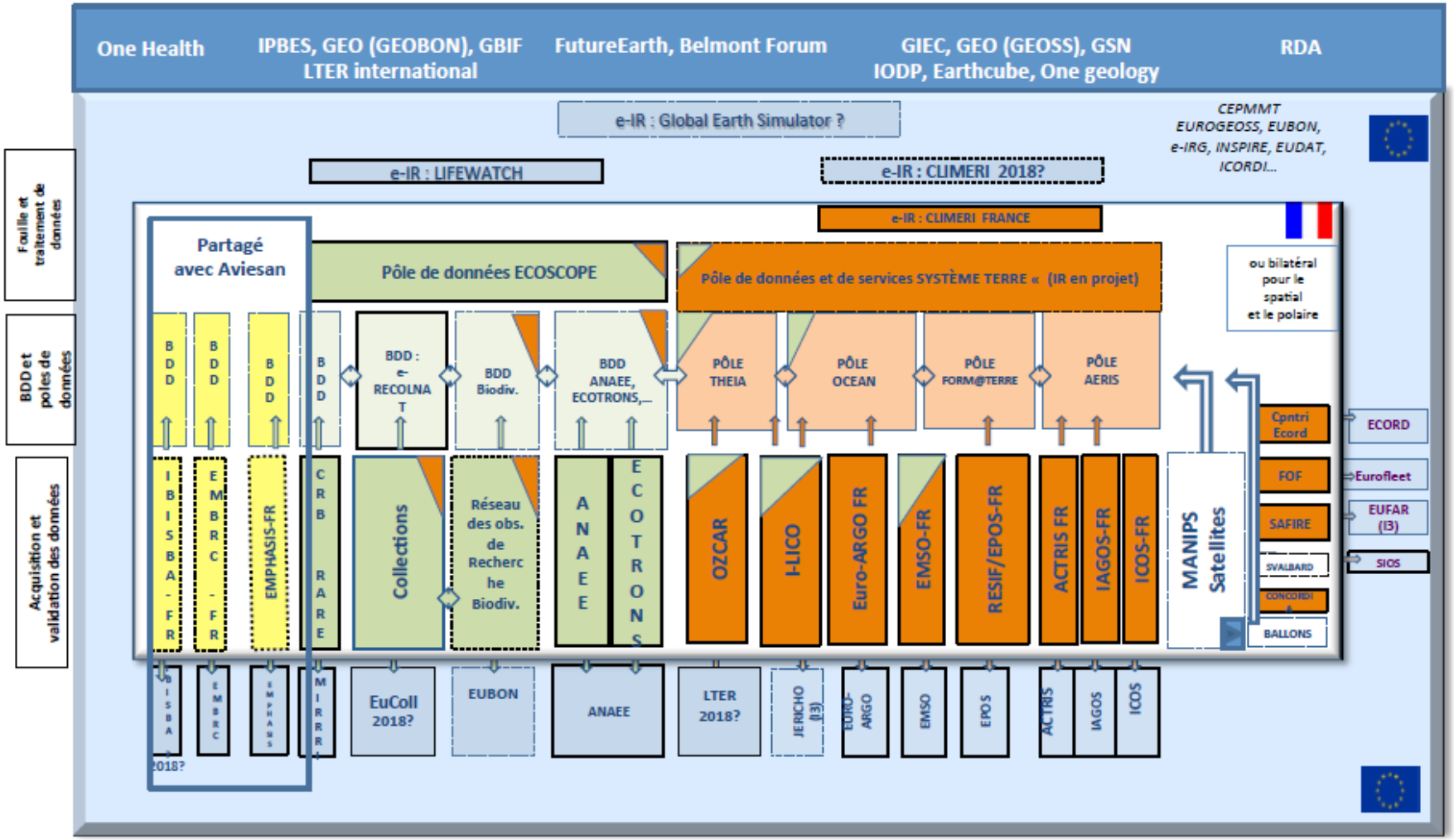
FAIR

data infrastructure: bridging islands




EUROPEAN OPEN SCIENCE CLOUD
BRINGING TOGETHER CURRENT AND FUTURE DATA INFRASTRUCTURES
A trusted, open environment for sharing scientific data
Open and seamless services to analyse and reuse research data
http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
Linking data
Connecting across borders and scientific disciplines
Connecting scientists globally
Improving science
Long term and sustainable

LES INFRASTRUCTURES DE RECHERCHE SYSTEME TERRE & ENVIRONNEMENT



DONNEES INFRASTRUCTURES

« Terre Vivante »

« Terre solide et fluide »

DONNEES SATELLITES

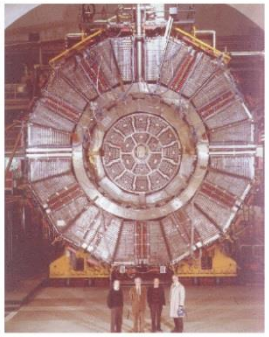
DONNEES CAMPAGNES

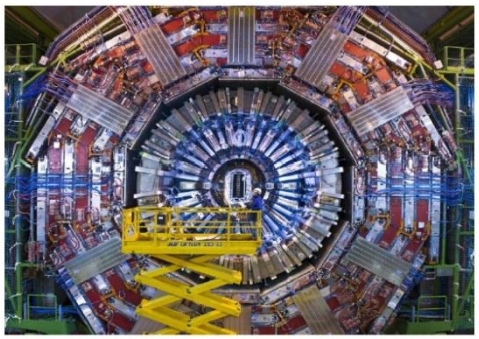
Légende — IR nationale | IR européenne | ESFRI

MINISTÈRE DE L'ÉDUCATION NATIONALE, DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Données en HEP

In Big Communities
In International Labs (CERN)





Past Century collaboration
~500 Scientists

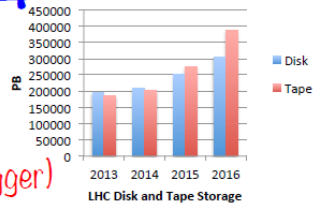
Today collaboration
~4000 Scientists

From all around the word

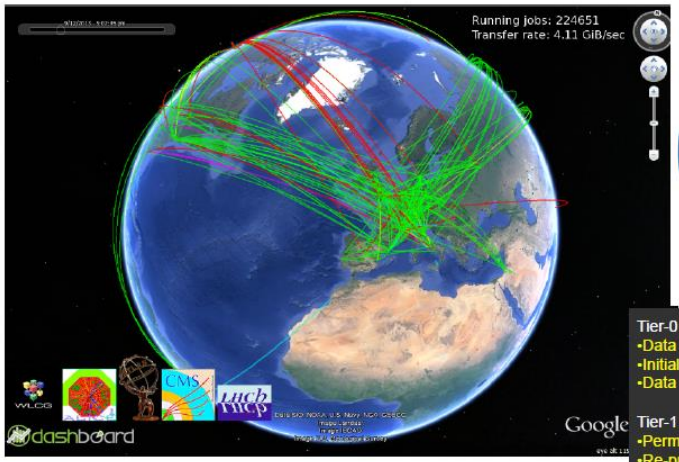
The (Big) DATA

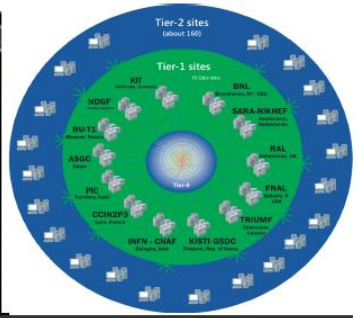
10^7 "sensors" produce 5 PByte/sec
Complexity reduced by a Data Model

Analytics in real time filters to 0.1–1 Gbyte/sec (Trigger)
Data + Replica move with a Data Management Policy
6 GB/s (600 TB/day)



Worldwide LHC Computing Grid





Tier-0 (CERN):
- Data recording
- Initial data reconstruction
- Data distribution

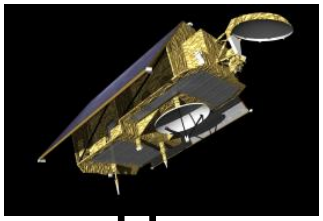
Tier-1 (12 centres):
- Permanent storage
- Re-processing
- Analysis

Tier-2 (68 Federations, ~140 centres):
- Simulation
- End-user analysis

- 525,000 cores
- 450 PB

Data Analytics exploit data by distributed computing infrastructure of half a million cores
An average of 40M jobs/month
produces
"Publication Data" that are openly Shared

Marcello Maggi
INFN Senior Researcher
Istituto Nazionale Fisica Nucleare
Bari-Italy



Copernicus Sentinel Data Policy

Sentinel Data Policy =
FREE and OPEN access

- Joint COM/ESA **Sentinel Data Policy Principles** have been prepared in 2009 - adopted by ESA MSs in Sep 2009
- **EU Delegated Act** on Copernicus Data and Information Policy has been adopted in 2013 (C(2013)4311, final)
- ESA got approval of updated **Sentinel Data Policy** from its Member States in Sep 2013. Main principles of Sentinel data policy:
  - **Open** access to Sentinel data by anybody and for any use
  - **Free** of charge data licenses
  - Restrictions possible due to technical limitations or security constraints

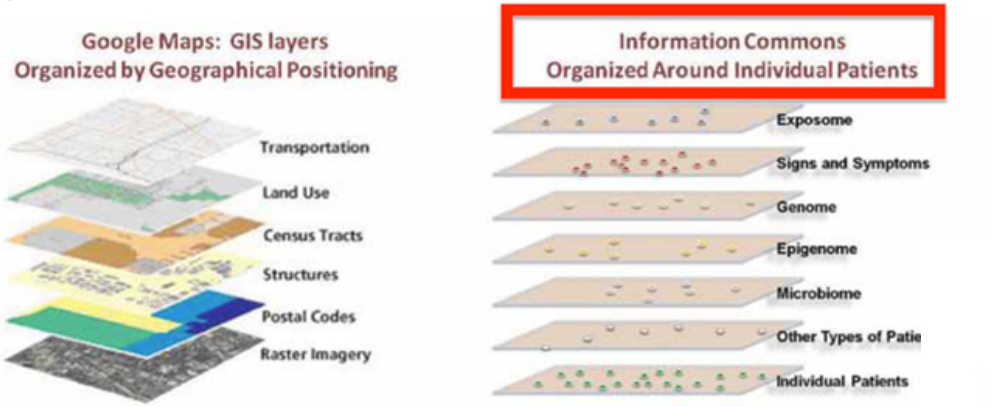
European Space Agency

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

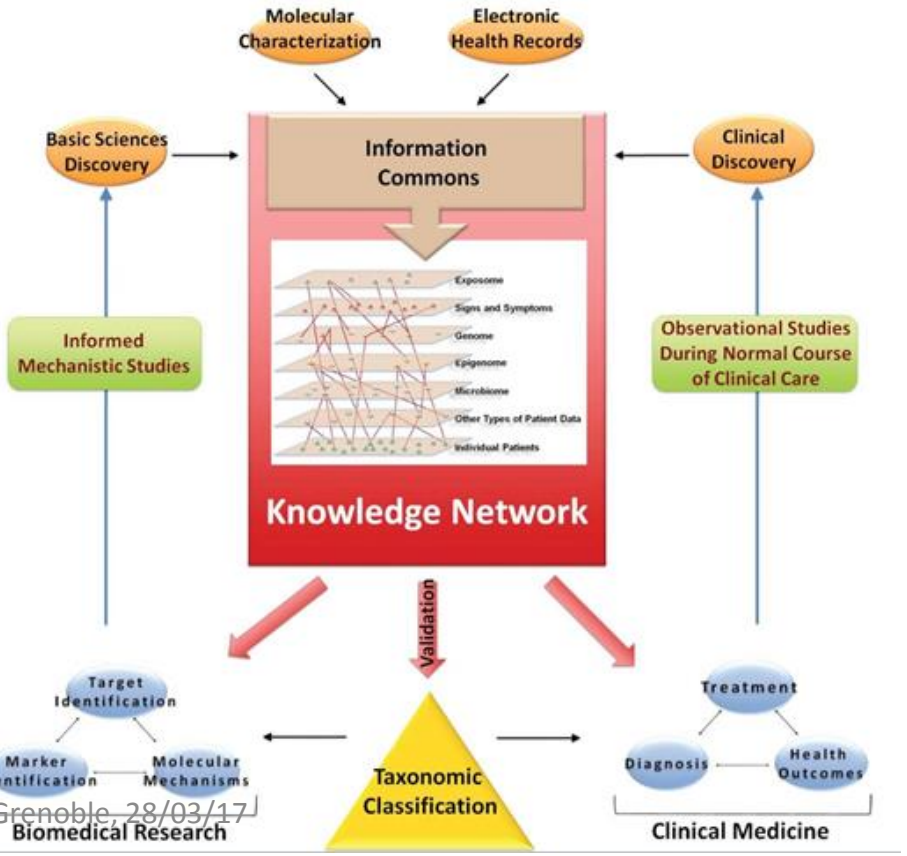
Les données de santé

Intégration des données pour une Médecine translationnelle, prédictive et personnalisée

Marc CUGGIA (MD,PhD)
Health Big Data team (LTSI) -
Clinical Investigation Center (CHU Rennes)
INSERM – Medical School
Université de Rennes 1 - BRITTANY

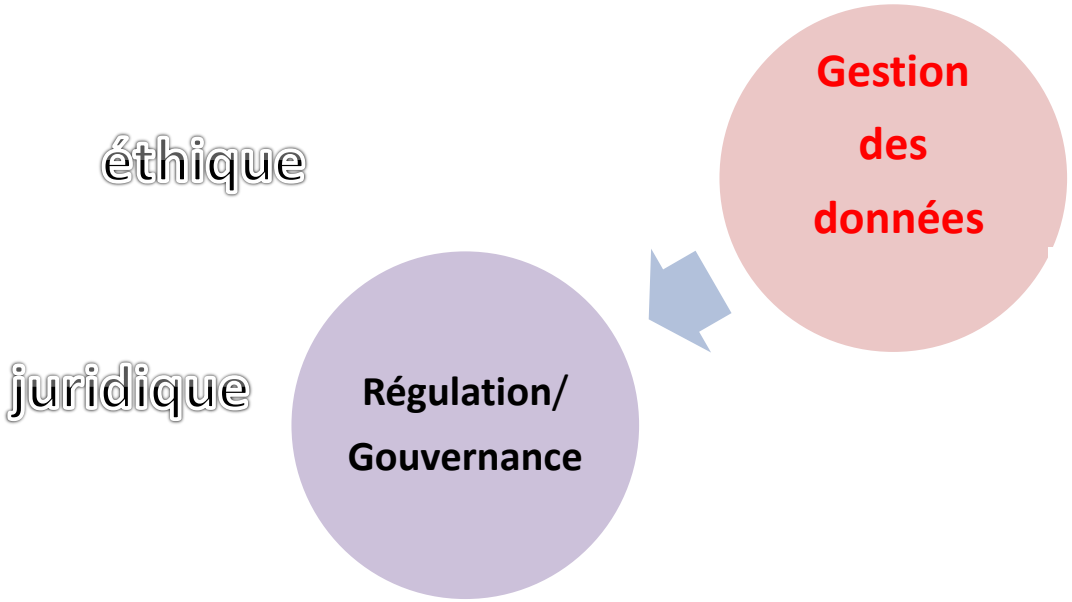


Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease
Report from National academy of science, USA, 2011



- Utilisation des données cliniques
- Développement d'une médecine personnalisée et prédictive

- relations gène/médicament, symptômes/maladies, risques environnementaux/expression des gènes



Politiques institutionnelles dont EC

Evolution du cadre juridique

LOI n° 2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public

- Principe de gratuité pour la réutilisation (sauf exceptions),
- Libre réutilisation par toute personne à d'autres fins que celle de la mission de service public, (abrogation exception)
- Incitation à la mise sous format ouvert et librement réutilisable,
- Possibilité de choisir une licence.

Décrets d'application en cours

LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique

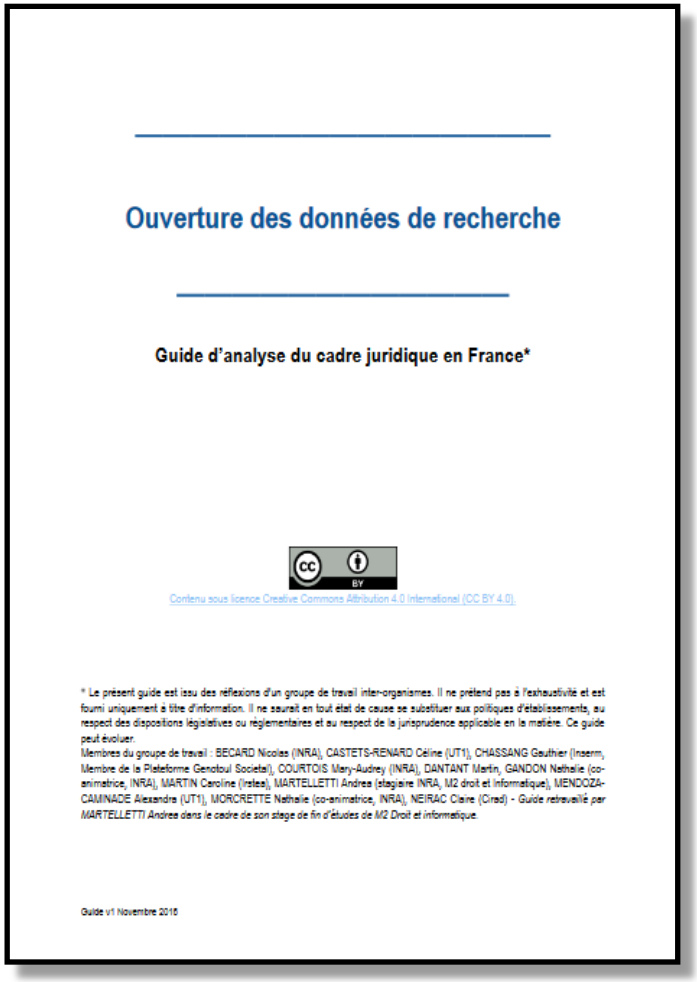
- Article 7 : réutilisation des données publiques (licences)
- Article 30 : permet la diffusion en libre accès des publications et des données
- Article 38 : facilite la fouille de texte et de données

A venir (2017): révision de la directive européenne sur les droits d'auteur et droits voisins

Cadre juridique national pour l’ ouverture des données

- Un contexte en évolution…

- Loi CADA, modifiée par la loi Valter
- Transposition de la directive PSI
- Loi Pour une république numérique
- A venir, modification de la directive européenne sur les droits d’auteurs et droits voisins



Ouverture des données de recherche

Guide d’analyse du cadre juridique en France*

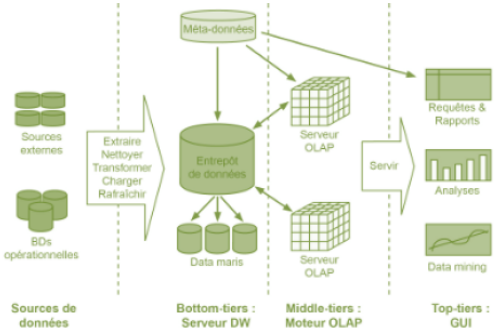
* Le présent guide est issu des réflexions d’un groupe de travail inter-organismes. Il ne prétend pas à l’exhaustivité et est fourni uniquement à titre d’information. Il ne saurait en tout état de cause se substituer aux politiques d’établissements, au respect des dispositions législatives ou réglementaires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

Guide v1 Novembre 2016

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

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 <u>interdictions totales</u> 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).

Extraits…

Nathalie Gandon, Nathalie Morcrette,Inra

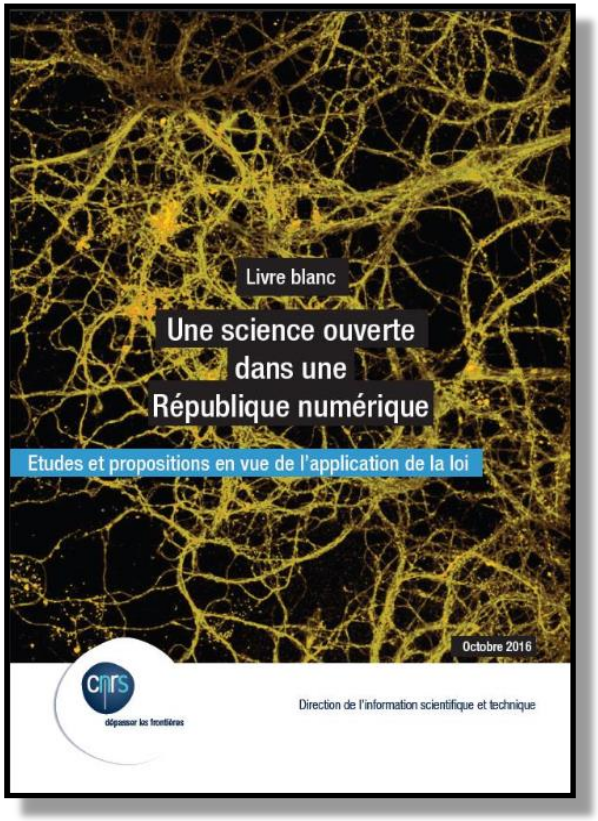
Exceptions : « données propriétaires »

- Données sont soumises au **droit d'auteur** : textes, plans, photographies, etc. et notamment les publications scientifiques.
  - ***Condition*** : **originalité** de la forme (pas de l'idée).
  - ***Conséquence*** : pour utiliser ces données, **l'accord de l'auteur** est indispensable (donc attention!! au text mining) sauf exception de courte citation .

  Le droit revient à l'auteur et non à l'établissement (sous réserve de conditions « d'autonomie »).

- Données organisées en bases de données : **Droit sui generis** qui peut s'appliquer sous réserve de la preuve d'un investissement substantiel (le plus souvent financier). Le droit revient à l'investisseur (le plus souvent l'établissement).

Cadre éthique et juridique de l'Open Data dans le contexte Open Science



Actualisation en cours du Livre Blanc « Loi pour une République numérique : Guide stratégique d'applications pour une Science ouverte »

Concourir à un meilleur partage des données ?

Mettre en conjonction …

- Pratiques des chercheurs : culture du partage/rewarding ?
- Pratiques des communautés : outils, logiciels, standards,
- Politique des infrastructures
- Politique des institutions : pérennité des moyens allouées aux infrastructures et à la gestion des données (y compris personnels d'accompagnement)
- Politique des financeurs : incitation/obligation DMP

BIBLIOTHÈQUE SCIENTIFIQUE NUMÉRIQUE
DIGITAL SCIENTIFIC LIBRARY

Propositions de BSN10 pour la mise en œuvre d'une politique nationale en matière de gestion des données de recherche

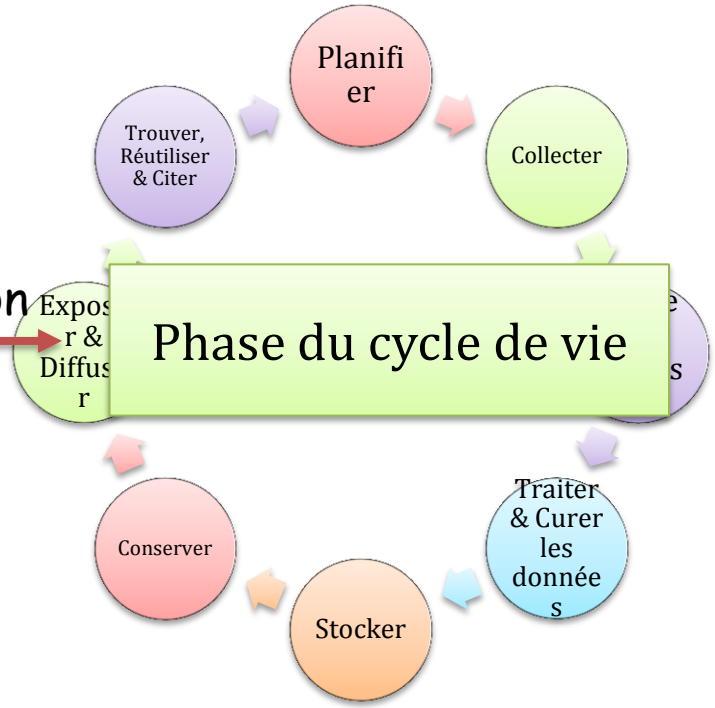
Travailler à …

- ❑ Définir des stratégies institutionnelles
- ❑ Mobiliser toutes les compétences : communauté scientifique, professionnels IST, informaticiens, archivistes, juristes,…
- ❑ Adapter les cadres éthique, juridique, économique.
- ❑ Adapter le dispositif d'évaluation des scientifiques et le cadre RH des fonctions d'appui
- ❑ Allouer les moyens de mise en œuvre de la politique en mobilisant des ressources pour : **dresser un état des lieux**, harmoniser les pratiques , clarifier les spécificités de chaque communauté, **produire un guide de bonnes pratiques**, rendre obligatoires les **plans de gestion de données (DMP)**, définir une stratégie de lobbying,…
- ❑ Développer des services de RDM pour les chercheurs

Actions clés

- ❑ Cartographier des acteurs dans les différentes composantes ESR (Cat OPIDoR)
- ❑ Mener des auditions : scientifiques pour clarifier les spécificités disciplinaires, groupe juridique Inra,…
- ❑ Participation au GT données du pilier INFRANUM

Cat OPIDoR



Service

Nom
URL
Typologie
Description
Discipline
Thématique
Localisation
Public cible
Contact (mail)
Champs spécifiques

Permet la réalisation

Offre

Est rattaché

Structure *

Nom
URL
Tutelles

*Institution, Organisme, Réseau, Fédération, Projet…

DMP OPIDoR

https://dmp.opidor.fr/

My plan (Horizon 2020 FAIR DMP)

Plan details | Initial DMP | Share | Export

1. Data summary (1 question, 0 answered) +

2. FAIR data (4 questions, 0 answered) −

In general terms, your research data should be 'FAIR' that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard or implementation-solution.

2.1 Making data findable, including provisions for metadata:

- Outline the discoverability of data (metadata provision)
- Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
- Outline naming conventions used
- Outline the approach towards search keyword
- Outline the approach for clear versioning
- Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

B I

Save

Guidance | Share note

CE Guidance −

The Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.

Digital Curation Centre guidance on Documentation +

Digital Curation Centre guidance on Metadata +

En guise de conclusion, quelques points de discussion…

- La plupart des enjeux du partage des données (et plus généralement de l'open science) ne sont pas techniques mais sociaux
- Au-delà du volume, apprendre à gérer la complexité et l'hétérogénéité des productions scientifiques
- La coopération entre les métiers est essentielle : apprendre le langage de l'autre !
- Les personnels de soutien (ing., inf., IST) ont un rôle à jouer dans le développement des pratiques de confiance
- Libérer du temps de recherche par des bonnes pratiques de RDM


Merci de votre attention
francis.andre@cnrs-dir.fr