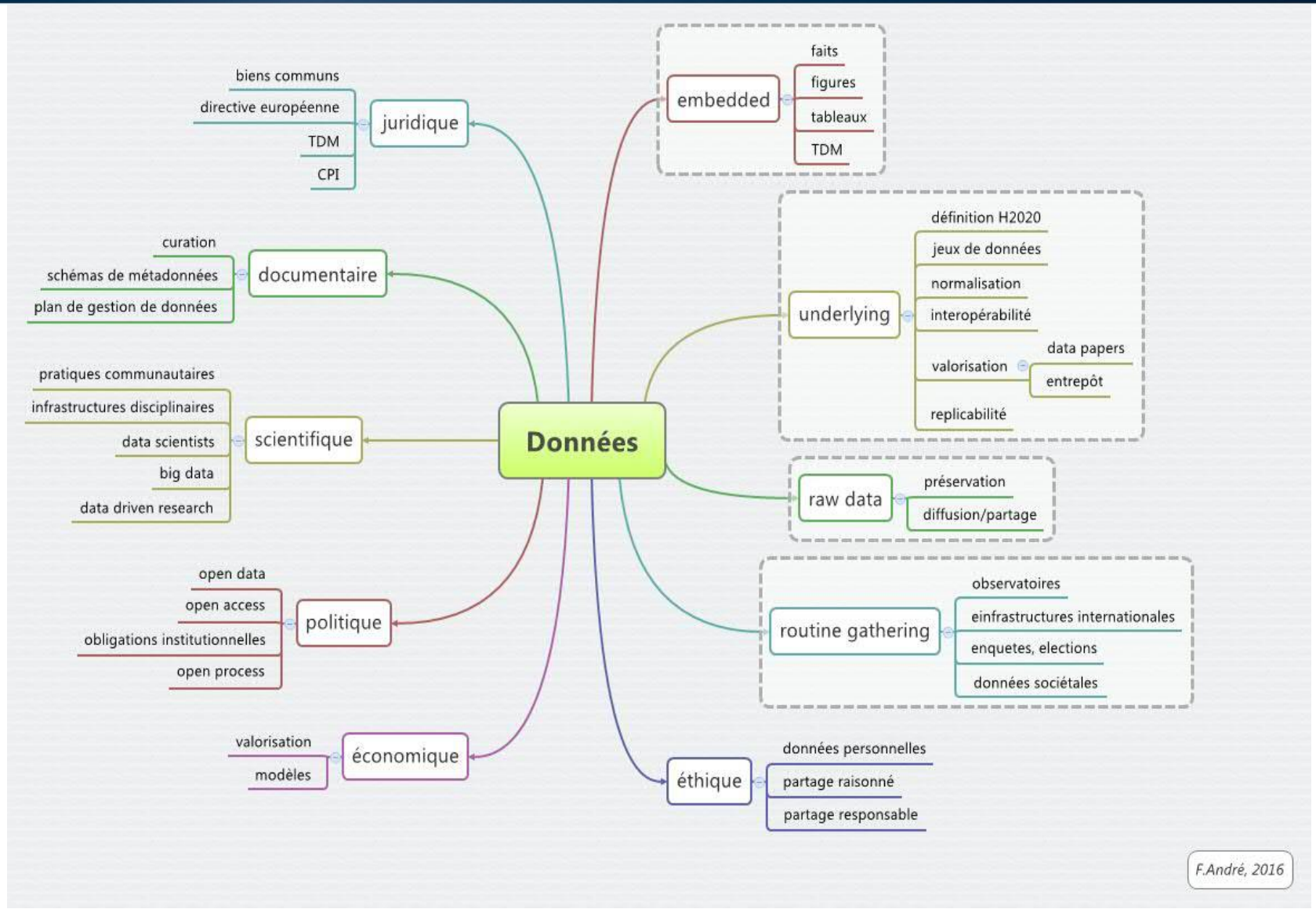
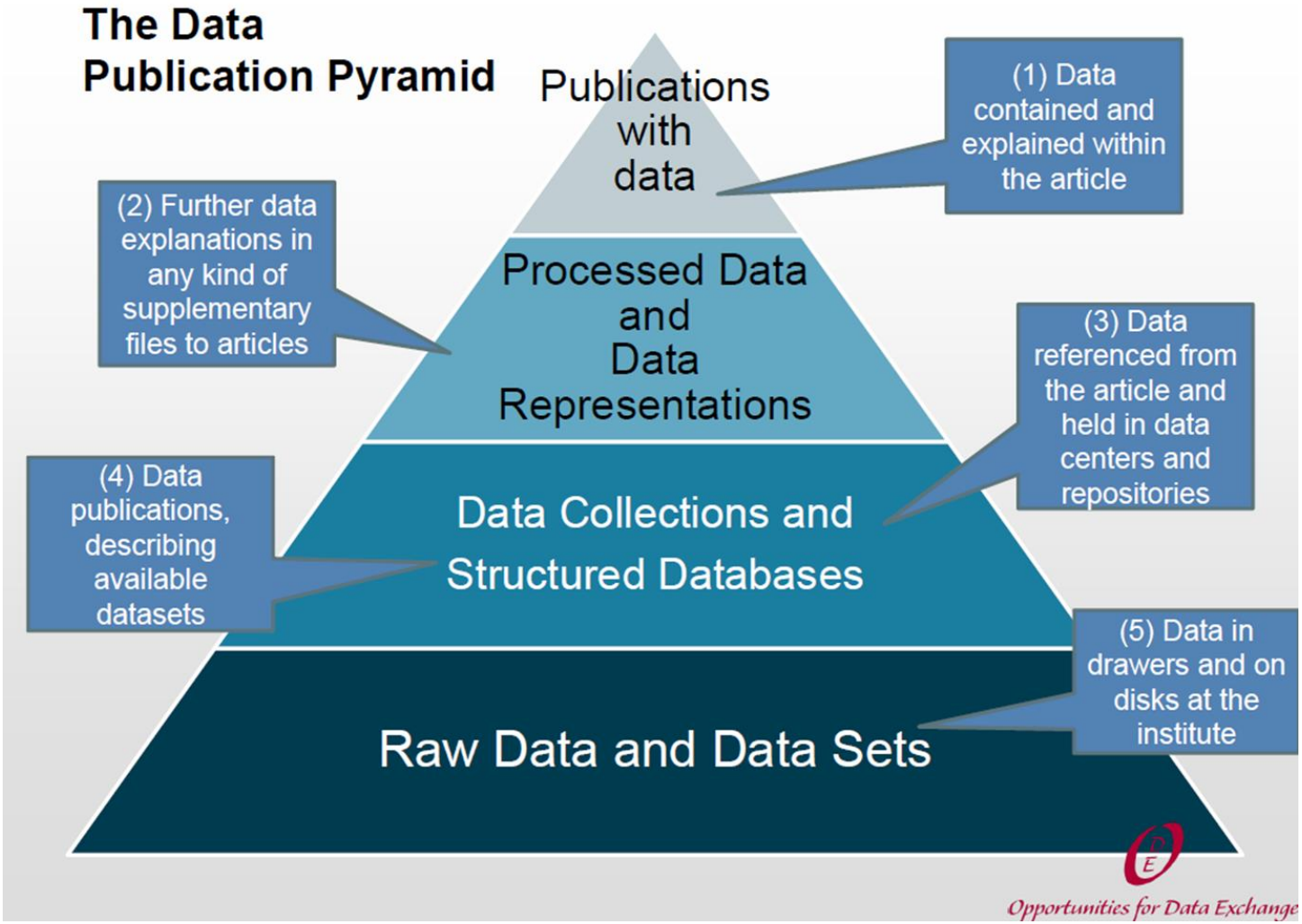
Le partage des données dans le contexte de la science ouverte

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

Partage des données : chantiers multiples



Localisation des données de recherche



Une définition 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"

Lien publication /données

Responsibility

Publications are arguments made by authors, and data are the evidence used to support the arguments.

C.L. Borgman (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. MIT Press

Gérer ? Partager ? Mais pourquoi ?

- Preuve scientifique
- Reproductibilité
- Réutilisabilité
- Circulation, accélérateur de l'innovation
- Efficience du processus de création de connaissance
- Enjeux sociétaux interdisciplinaires
- Enregistrement de la connaissance
- …

Pourquoi parler de science ouverte ?

Science du 21ème siècle : plus…

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

Open Research

Science 2.0

eScience

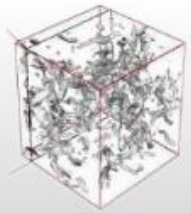
OPEN SCIENCE

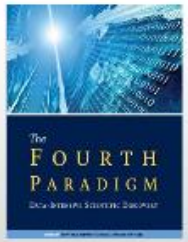
Contexte: accroissement de la production scientifique, du nombre de chercheurs, nouvelle façon de faire de la science, guidée par les données massives, importance des défis sociétaux

Faire de la science…

La “science des données”, 4e paradigme de la découverte scientifique

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$





Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago	Last few hundred years	Last few decades	Today and the Future
Description of natural phenomena	*Newton’s laws, Maxwell’s equations…*	*Simulation of complex phenomena*	*Unify theory, experiment and simulation with large multidisciplinary Data* *Using data exploration and data mining (from instruments, sensors, humans…)* *Distributed Communities*

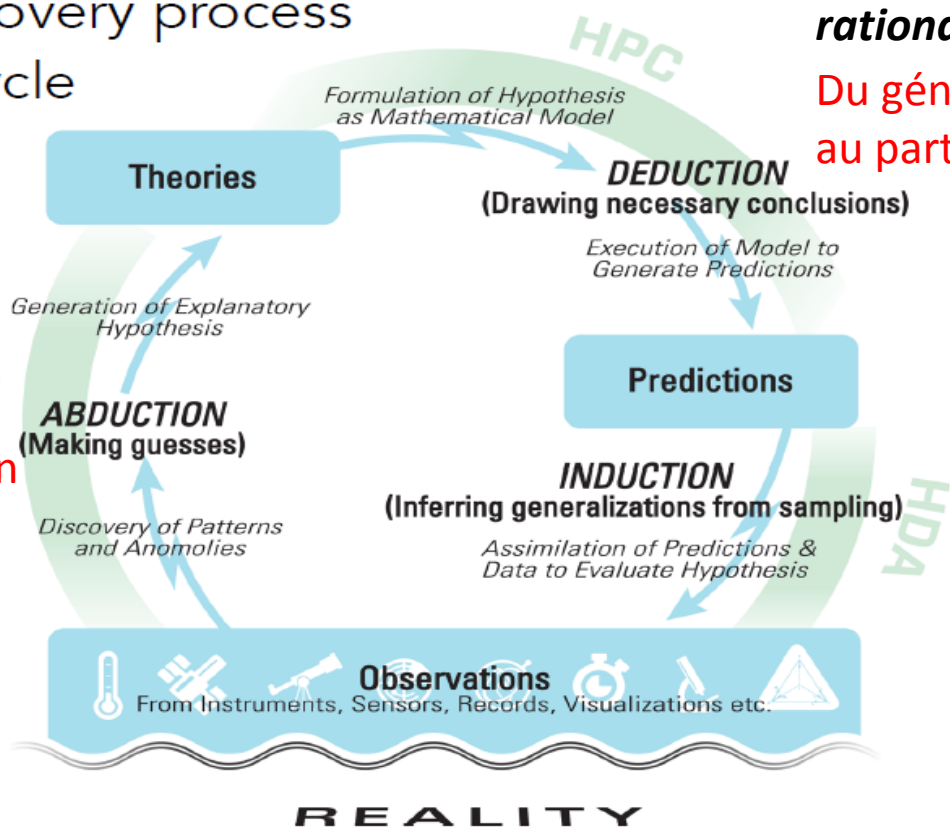
Crédits: Dennis Gannon

Inria

Données et création de connaissance

Context: a new paradigm

The scientific discovery process
= the inference cycle



rationalisme

Du général
au particulier

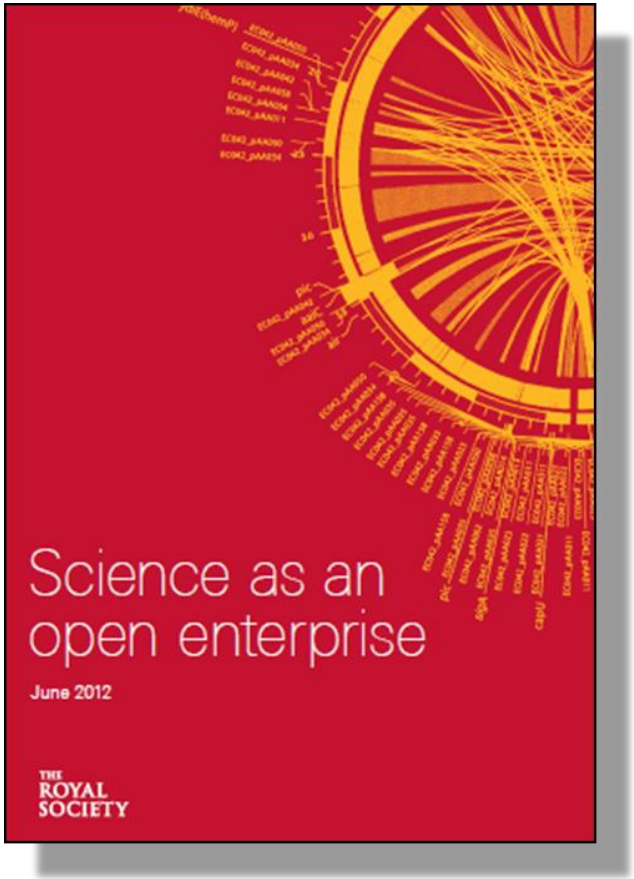
pragmatisme

Formulation de la
meilleure explication

empirisme

Du particulier
au général

Conseils de lecture…

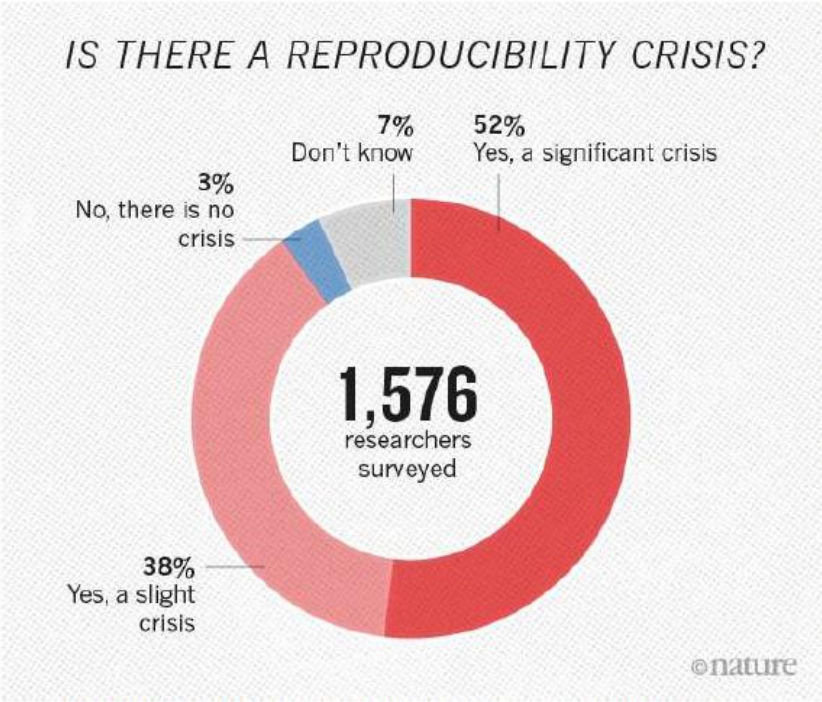


https://www.innovationpolicyplatform.org/content/open-science

https://royalsociety.org/topics-policy/projects/science-publicenterprise/report/

Reproductibilité ?

1 Issue: The Reproducibility Crisis



Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

https://www.slideshare.net/AustralianNationalDataService/research-data-management-in-practice-ria-data-management-workshop-brisbane-2017

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

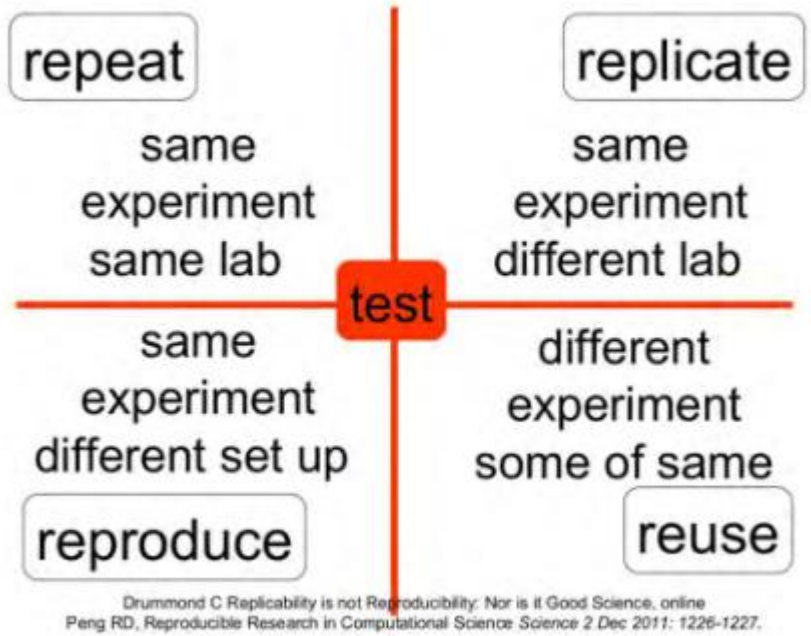
- **A methodological crisis in science**
- the phrase was coined in the early 2010s as part of a growing awareness of the problem
- 2016: poll of 1,500 scientists
- 70% of them had failed to reproduce at least one other scientist's experiment
- results of many scientific studies are difficult or impossible to replicate on subsequent investigation

Funding agencies have already taken notice

La reproductibilité, exemple de la bio-informatique

- Reproductibilité *computationnelle*
- Nombre croissant de résultats scientifiques non reproductibles
  - Y compris dans les revues à fort facteur d'impact
  - Pas (toujours) volontairement
- Nombreux domaines concernés
  - Certains plus critiques que d'autres…
- Enjeux économique majeur
  - Non reproductibilité des études pré-cliniques évalué à >$10 milliards annuel pour les USA
- Devient une obligation contractuelle
  - Projets NSF, certains editeurs

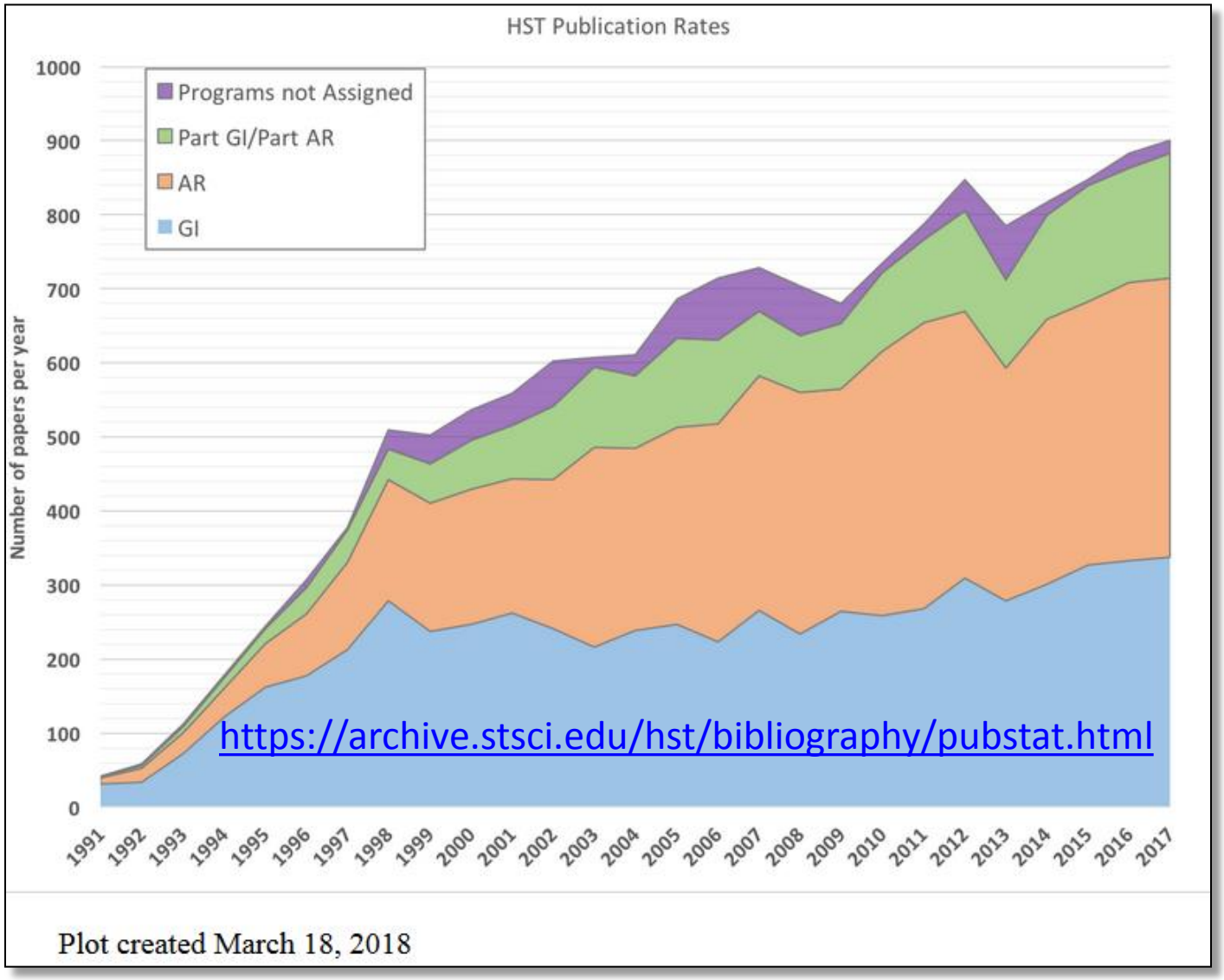
Sarah Cohen-Boulakia, Univ Paris Sud



Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

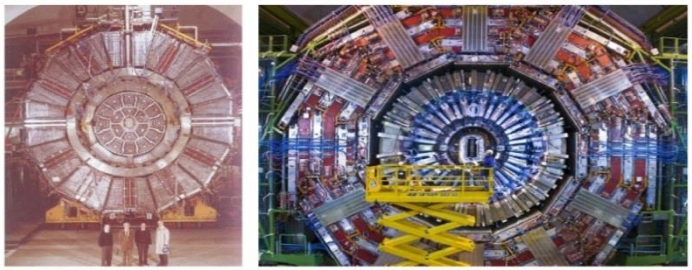
Les données, mais aussi processus, workflows, environnement d'exécution,…

Réutilisation des données du télescope Hubble



Partager les données en HEP

In Big Communities
In International Labs (CERN)



Past Century collaboration
~500 Scientists

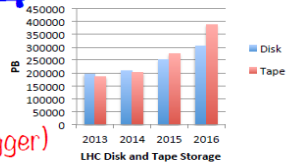
Today collaboration
~4000 Scientists

From all around the word

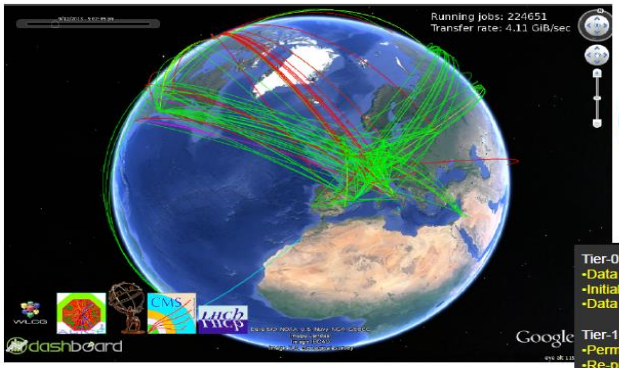
The (Big) DATA

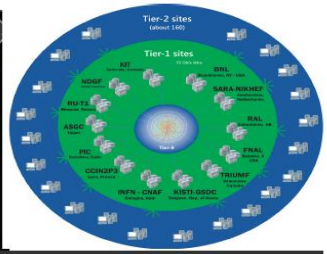
10^7 "sensors" produce 5 PByte/sec
Complexity reduced by a Data Model

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

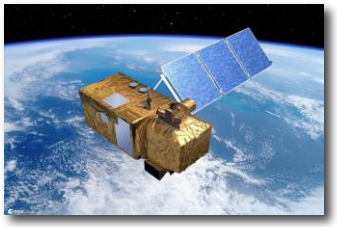
Data Analytics exploit data by distributed computing infrastructure of half a million cores
An average of 40M jobs/month

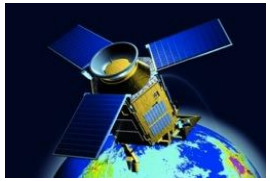
produces

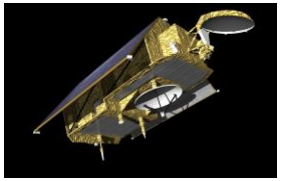
"Publication Data" that are openly Shared

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

Partager les données : Copernicus







Copernicus Sentinel Data Policy

Sentinel Data Policy =
FREE and OPEN access

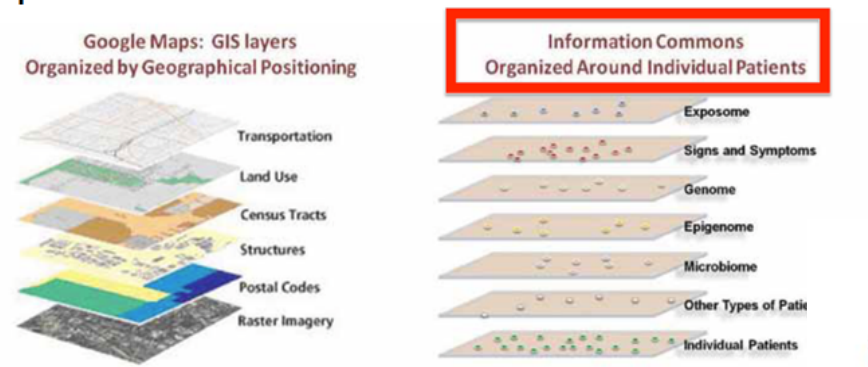
- Joint COM/ESA **Sentinel Data Policy Principles** have been prepared in 2009 - adopted by ESA MSs in Sep 2009
- **EU Delegated Act** on Copernicus Data and Information Policy has been adopted in 2013 (C(2013)4311, final)
- ESA got approval of updated **Sentinel Data Policy** from its Member States 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

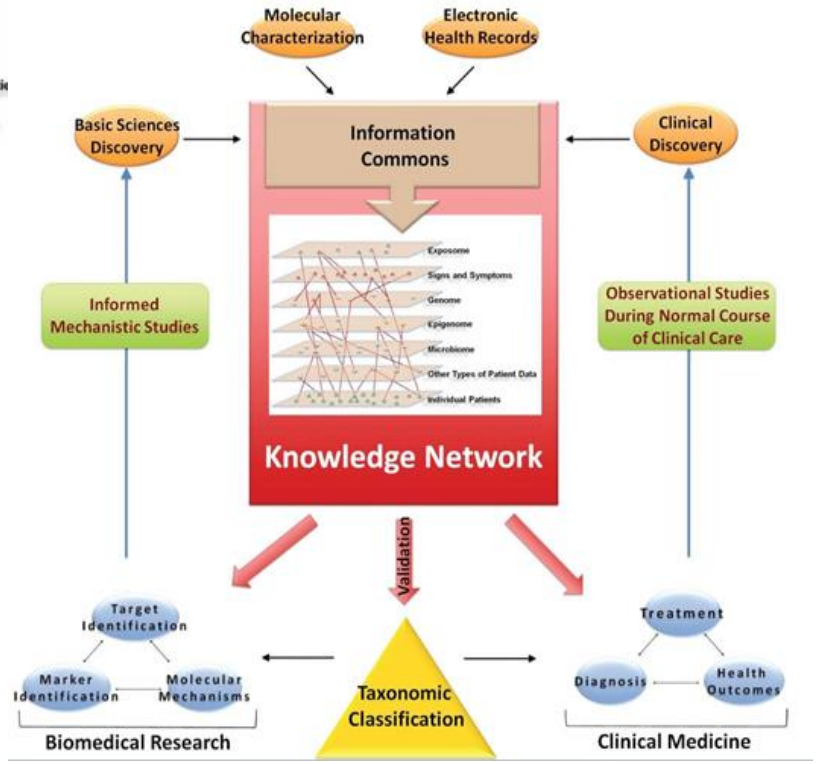
Partager les données de santé

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

La coordination européenne en sciences de l'environnement

ENVRIPLUS – EC FUNDED CLUSTER PROJECT OF ENVIRONMENTAL RESEARCH INFRASTRUCTURES

- 15 M € budget for 2015-2019
- 37 beneficiaries
- 19 Work Packages organized in 6 Themes
- 22 Research Infrastructures from 4 Environmental Domains

*Aim is to create **more coherent, interdisciplinary and interoperable** cluster of Environmental Research Infrastructures in Europe*

Les challenges du partage en sciences de l'environnement

Quality challenge:
Quality control for large quantity and nearly real time sensor data

Collaboration challenge in **documenting data processing workflows and sharing among communities.**

Identification/ Citation/ challenge
Identifying and **citing** data objects, and in **publishing** data

ENVRI plus

Execution challenge in **executing** applications on **distributed computing infrastructures**

Access challenge:
accessing data from **different sources**

Processing challenge in **combining** different data **processing models**

Principes FAIR

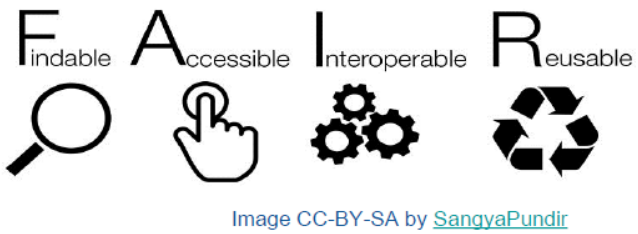
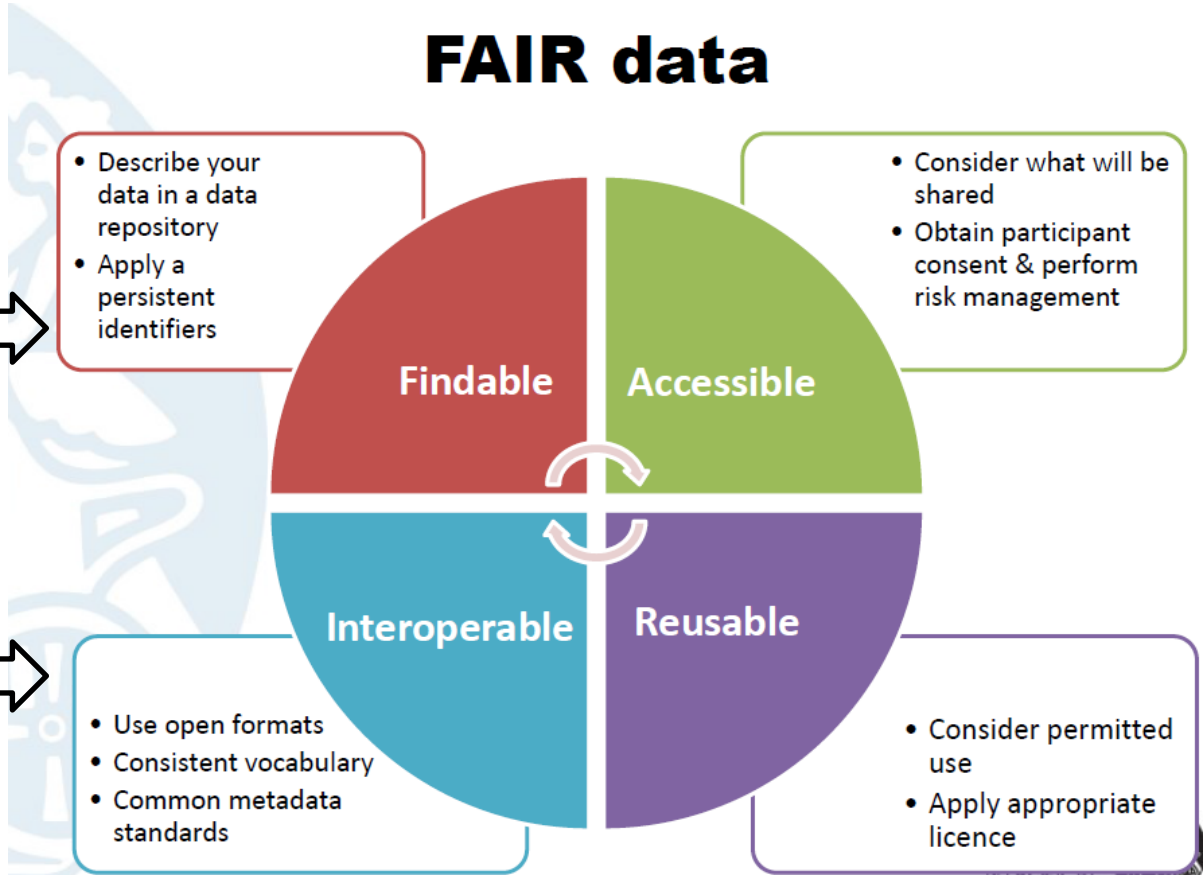


Image CC-BY-SA by SangyaPundir

Presque simple…

De vrais challenges !



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

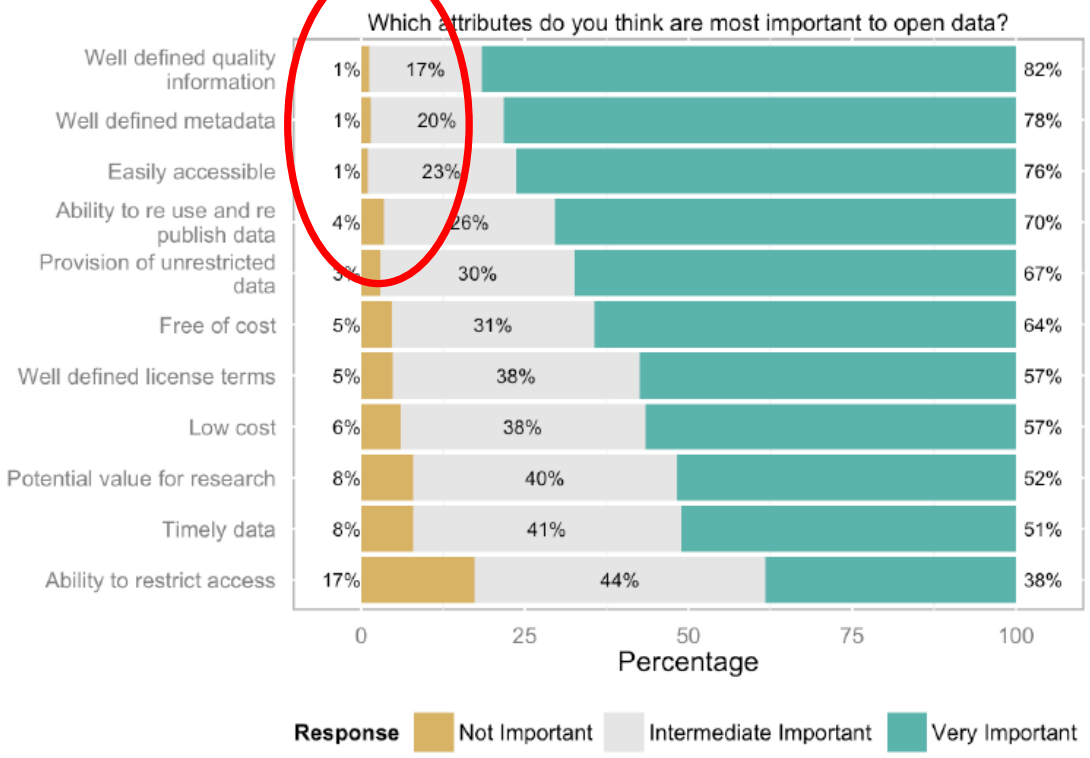
Où se niche la qualité des (méta)données ?

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

Qualité des données…

What properties do they expect for open data?

Belmont Forum survey 2016



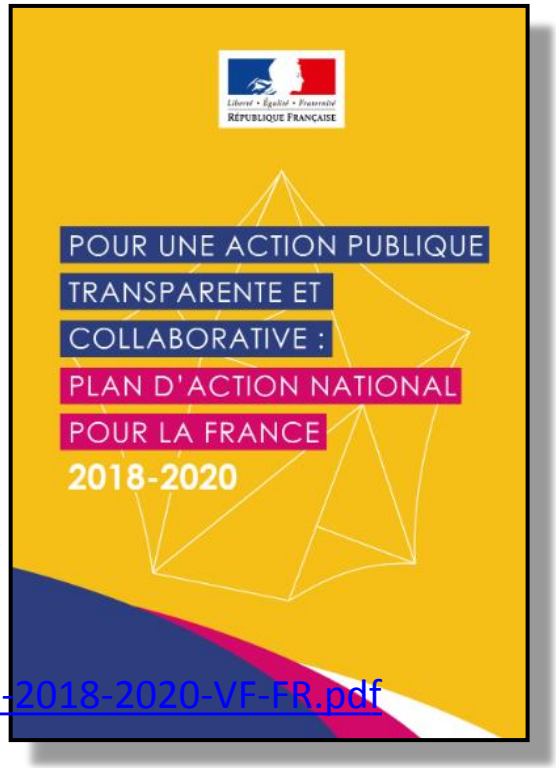
10

L'ouverture des données, l'axe politique

- As open as possible, as closed as necessary
- Open gouvernement partnership
  - Ouverture des données publiques
- Cohérence internationale :
  - G8, Europe, OCDE, >>> RDA,…
- Cohérence nationale :
  - TGIR/IR, CoSO, >>>RDAFrance…

ENGAGEMENT 18

Construire un écosystème de la « science ouverte »



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

Initiative nationale : PNSO

- Plan national pour la Science Ouverte, 4 juillet 2018
- Comité pour la Science Ouverte : collège Données

- **« les données de la recherche sont la matière première de la connaissance. Les partager c'est ouvrir de nouvelles perspectives scientifiques »**

Initiative nationale : PNSO

DEUXIÈME AXE : **STRUCTURER ET OUVRIR LES DONNÉES DE LA RECHERCHE**

ANR Flash : adoption des principes FAIR et ouverture des données

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.

- Généralisation du DMP dans les appels à projets (ANR en 2019)
- Soutien à la Research Data Alliance

Initiative nationale : PNSO

TROISIÈME AXE : **S'INSCRIRE DANS UNE DYNAMIQUE DURABLE, EUROPÉENNE ET INTERNATIONALE**

Créer un label Science Ouverte pour les écoles doctorales

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*.

- Généralisation des compétences de la Science Ouverte
- Inscrire les actions du CoSO dans le paysage international : collège Europe et International
- Participer à EOSC

Open Science Policy Platform

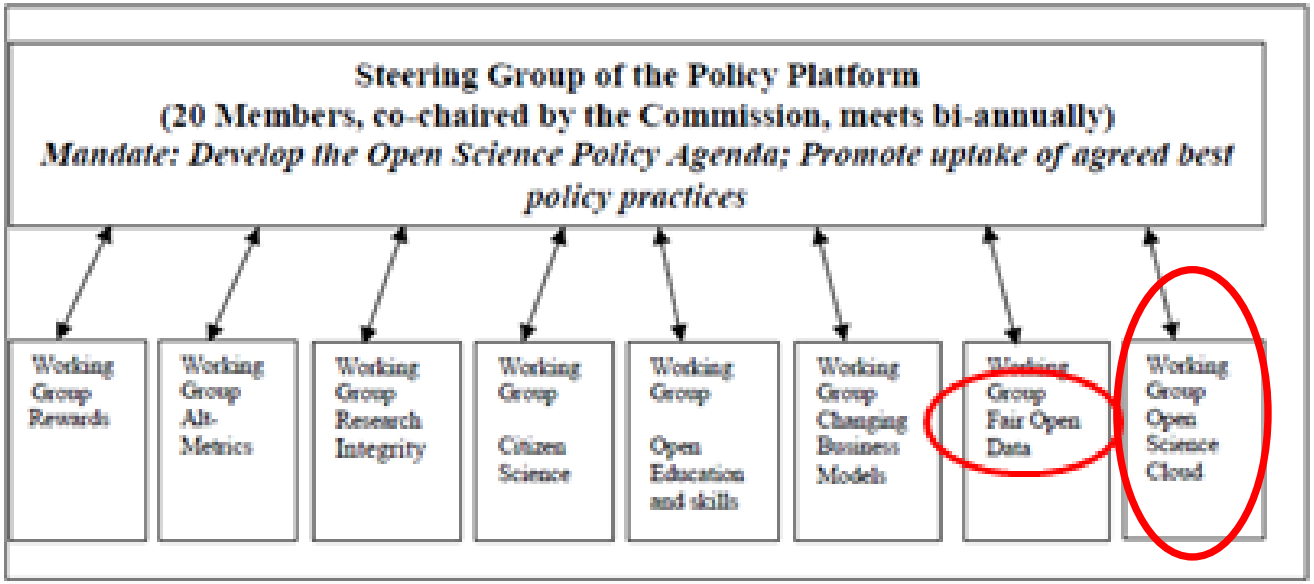
2016

European Commission

DIRECTORATE-GENERAL FOR RESEARCH AND INNOVATION (RTD)

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

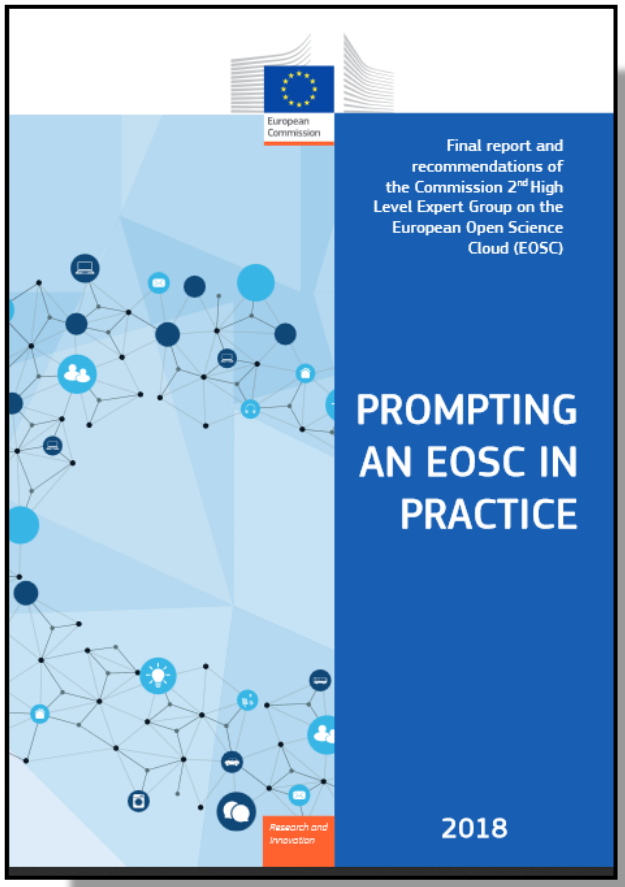
http://ec.europa.eu/research/openscience/pdf/ospp_nominated_members.pdf#view=fit&pagemode=none



FAIR : Findable, Accessible, Interoperable, Reusable

De la lecture …

http://www.donneesdelarecherche.fr/spip.php?article1610



http://www.donneesdelarecherche.fr/spip.php?article1609

Objet FAIR ?

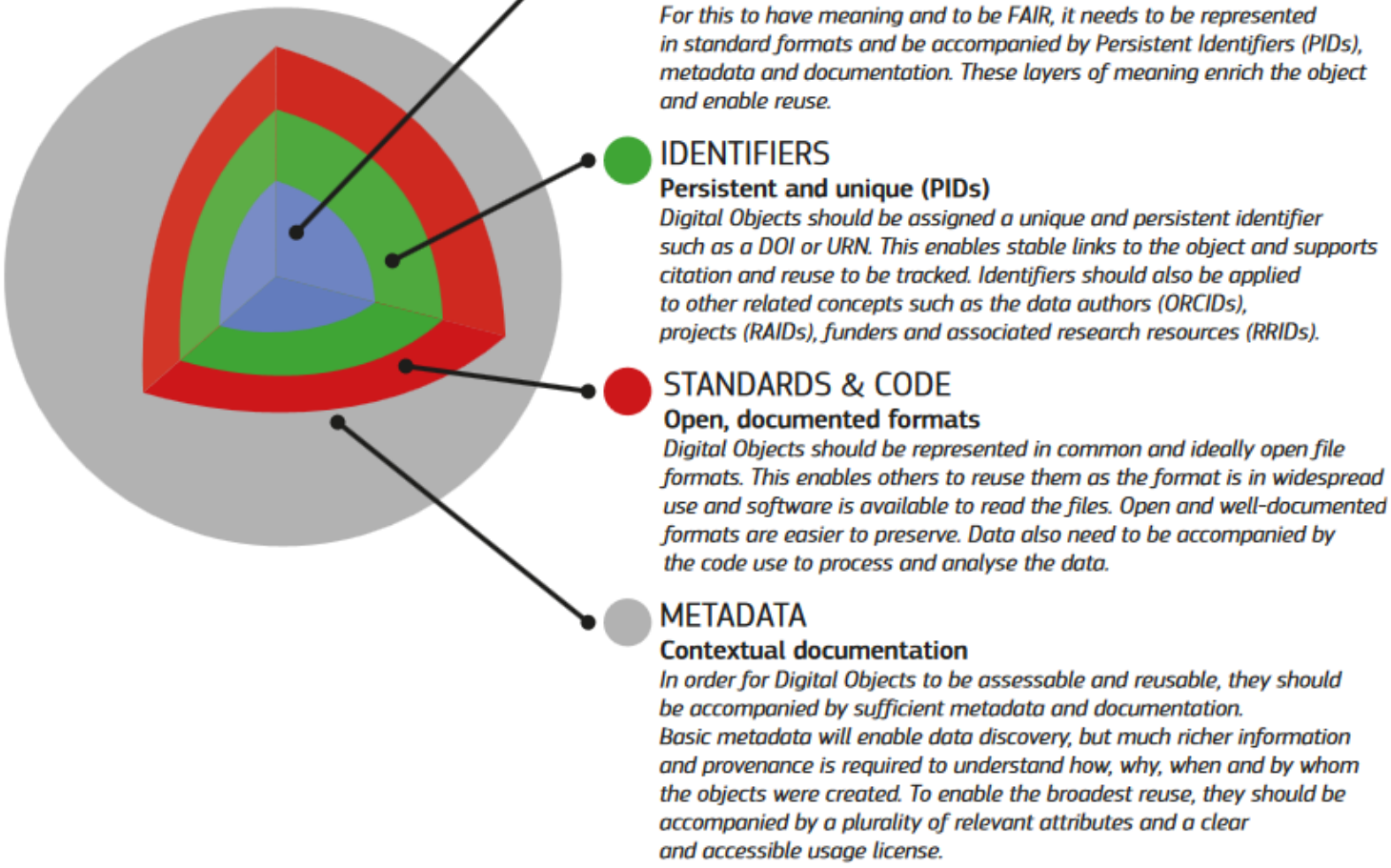


Figure 8. A model for FAIR Digital Objects

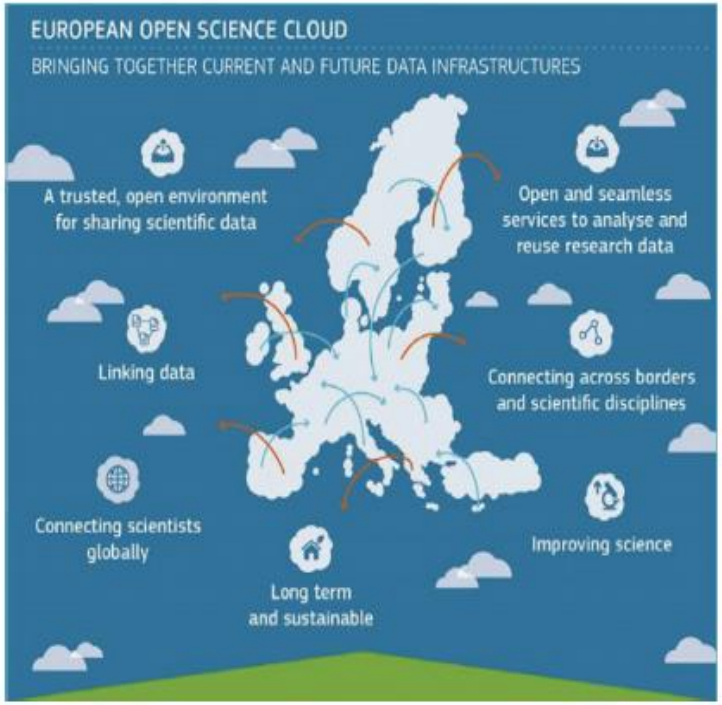
European Open Science Cloud

European Commission

"Europe's final transition must be one from fragmented data sets to an integrated European Open Science Cloud. By 2020, we want all European researchers to be able to deposit, access and analyse European scientific data through a European Open Science Cloud.."

Speech by Commissioner Carlos Moedas in Amsterdam, NL: "Open science: share and succeed", 4 April 2016

- EOSC will provide 1.7m EU researchers an environment with free, open services for data storage, management, analysis and re-use across disciplines
- EOSC will join existing and emerging horizontal and thematic data infrastructures, bridging todays fragmentation and ad-hoc solutions
- EOSC will add value (scale, data-driven science, inter-disciplinarity, faster innovation) and leverage past infrastructure investment (10b per year by MS, two decades EU investment)



EOSC, une initiative préparée de longue date

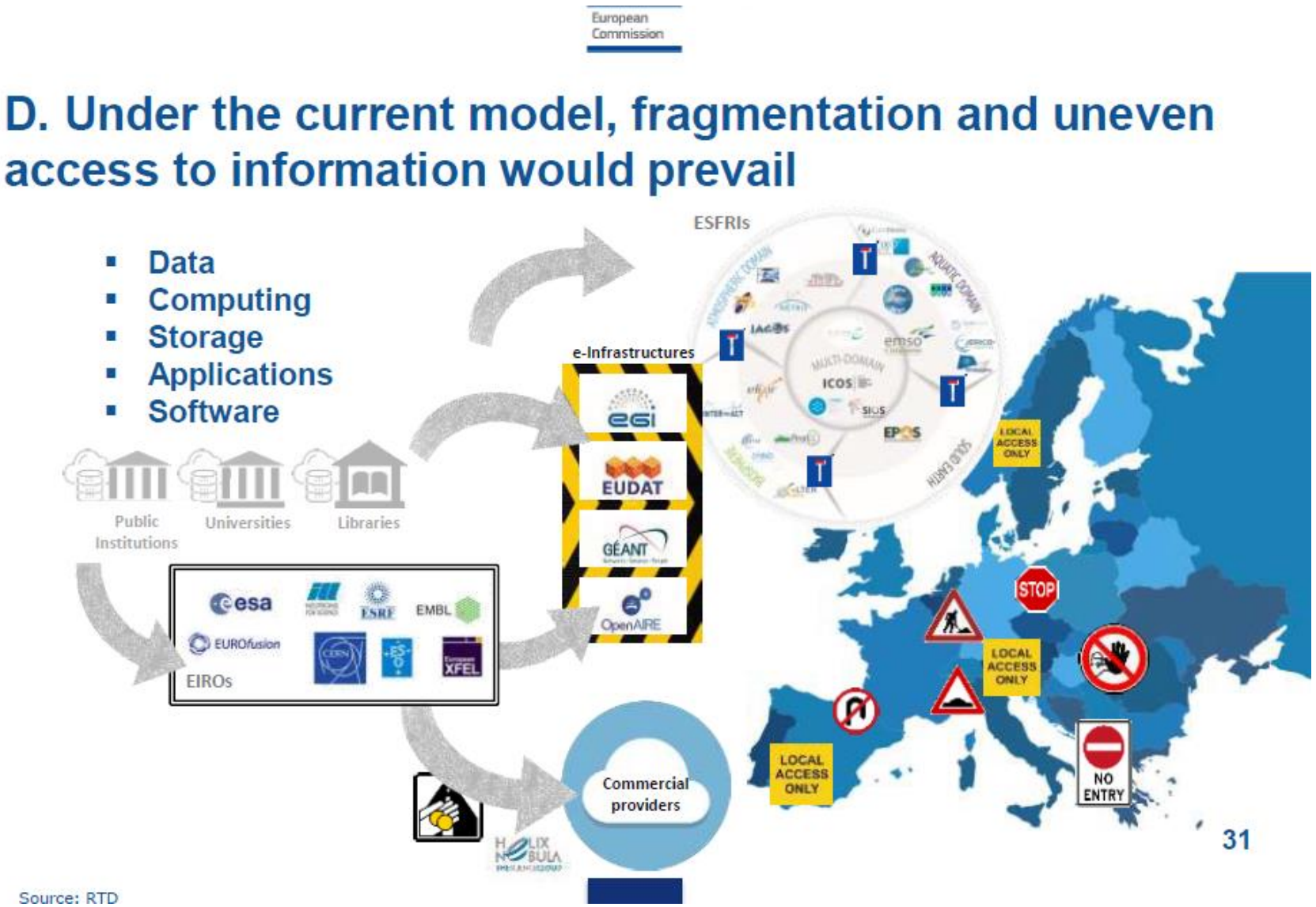
Actions from WP 2016-2017 of the Research Infrastructure paving the way for the establishment of the EOSC

- Key projects have been launched in January 2018

- Joining other existing relevant initiatives/projects

EOSC pilot
The European Open Science Cloud for Research Pilot Project
eInfra Central
GÉANT
RDA
RESEARCH DATA ALLIANCE
European Commission

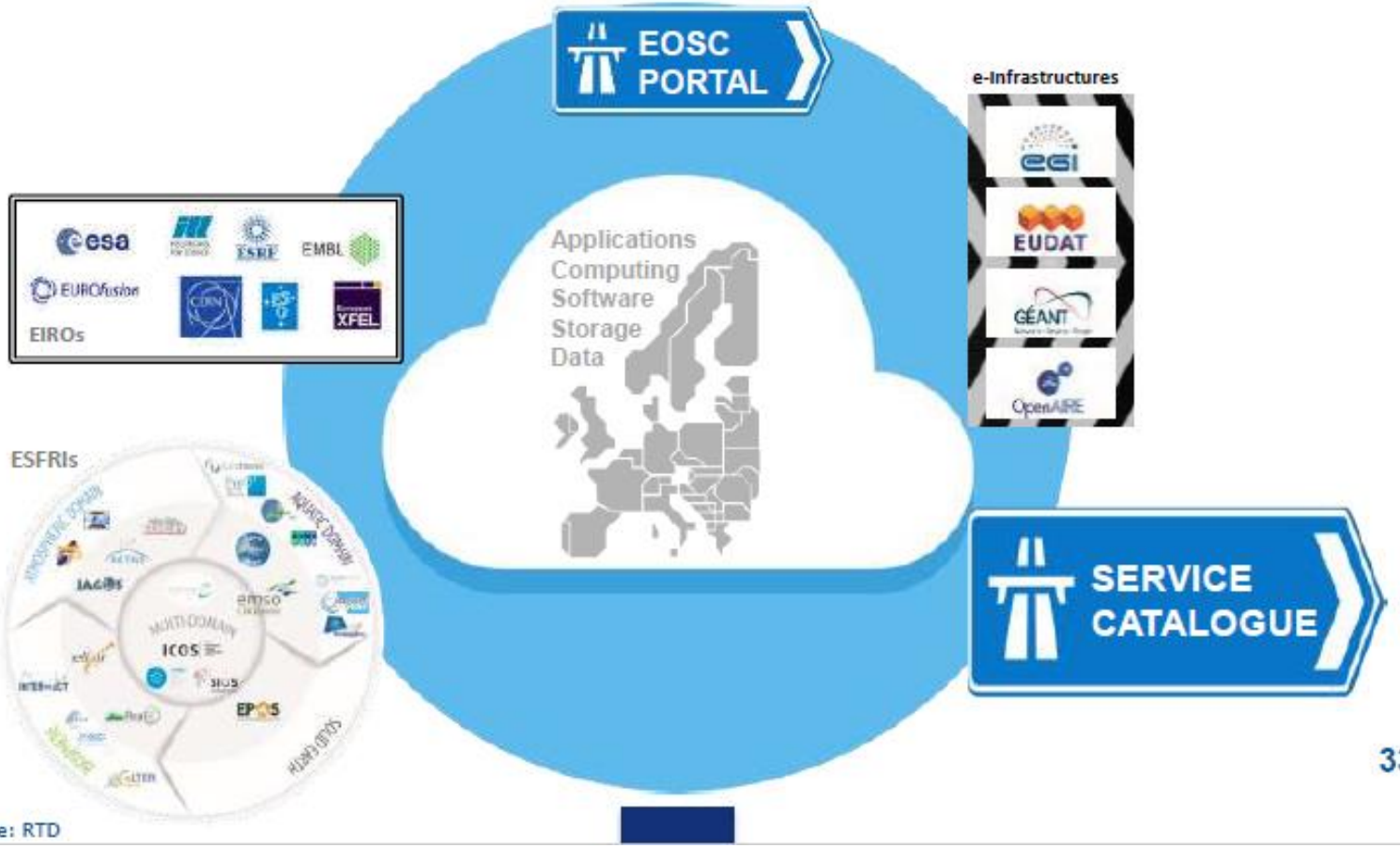
Fonctionnement actuel



Modèle fédératif

European Commission

D. Under the federated model, access to data would be universal, building on a strong legacy



Source: RTD

European Commission

E. The 6 lines of action of the EOSC model

a. Architecture	Architecture of the federated infrastructures as the solution to the current fragmentation in research data infrastructures which are insufficiently interoperable.
b. Data	FAIR data management and tools. A common data language to ensure data stewardship across borders/disciplines based on FAIR principles.
c. Services	Available services from a user perspective. A rich environment offering a wide range of services covering the needs of the users.
d. Access & Interface	Mechanisms/interfaces for accessing EOSC. A simple way for dealing with open data obligations or accessing research data across different disciplines.
e. Rules	Rules of participation for different EOSC actors. An opportunity to comply with existing legal and technical frameworks and increase legal certainty & trust.
f. Governance	Governance of the EOSC, aiming at ensuring EU leadership in data-driven science but requiring new governance frameworks.

Source: RTD

Des actions coordonnées

European Commission

(b) EOSC model/Data: Main actions

1. Work towards a **FAIR Data Action Plan**
2. Propose a **European Framework for FAIR Research Data** in line with the existing European Interoperability Framework
3. Analyse the **legal landscape** concerning data reusability
4. Develop a **FAIR Data accreditation/certification** scheme
5. Establish a cross-disciplinary **Persistent Unique Identifier** policy
6. Develop a **Catalogue of data standards**

RECHERCHE DE CONSENSUS

Building on: "FREYA", "OpenAIRE", "RDA Europe", "eInfraCentral", "EOSC Hub" projects and other EOSC-related projects

47

Et la suite !

European Commission

(b) EOSC model/Data: Milestones (**SWD** and interim steps)

Q3 2018	**FAIR Data Action Plan** including proposal for a European Framework for FAIR Research Data
Q2 2019	**European Framework for FAIR Research Data** FAIR Data Legal landscaping
Q4 2019	**FAIR Data accreditation/certification scheme** **FAIR Data Persistent Unique Identifier policy** **Initial Catalogue of data standards**
Q2 2020	**Catalogue of data standards**

Source: RTD

Des initiatives bottom-up : Research Data Alliance

L'ALLIANCE POUR LES DONNEES DE RECHERCHE

www.rd-alliance.org

Construire les ponts sociaux et techniques permettant un libre partage des données

28 REALISATIONS PHARES

dont 4 spécifications techniques ICT

75 CAS D'UTILISATION

dans différents domaines, organisations et pays

94 GROUPES TRAITANT DE L'INTEROPERABILITE GLOBALE DES DONNEES

dont 32 GROUPES DE TRAVAIL et 61 GROUPES D'INTERET

7 372 MEMBRES INDIVIDUELS issus de 137 PAYS

67% recherche & académique
14% administration publique
11% entreprise & industrie

48 ORGANISATIONS MEMBRES ET 8 MEMBRES AFFILIES

Vision

Les acteurs de la recherche et de l'innovation partagent librement les données, quels que soient les technologies, les disciplines et les pays, afin de répondre aux grands défis de société.

Mission

La RDA construit des passerelles à la fois sociales et techniques pour permettre un libre partage des données.

Nœud national

RESEARCH DATA ALLIANCE EUROPE

FR NATIONAL NODE

Quelques invariants

- Les données (FAIR data) sont un élément essentiel de la ***politique de Science Ouverte***
- 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 activités (***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 de données*** et services de partage.

« Take home messages »

- Ma recherche est-elle reproductible ?
- Mes workflows sont-ils robustes, documentés ?
- Mes logiciels sont-ils documentés, ouverts ?
- Mes données sont-elles ouvertes, pérennes, curées ?
- Mes données sont-elles diffusables, réutilisables ?

- A suivre ! FAIR en pratique au cours du cycle de vie des données

Et d'autres sujets : rewarding, évaluation,
données et sciences citoyennes, qualité et certification des entrepôts, infrastructures,…


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