



Participer à l'organisation du management des données de la recherche : gestion de contenu et documentation des données

6-8 juil. 2016 Paris (France)



www.cnrs.fr

Les données de la recherche dans le contexte de l'Open Science

Francis André CNRS/DIST

francis.andre@cnrs-dir.fr



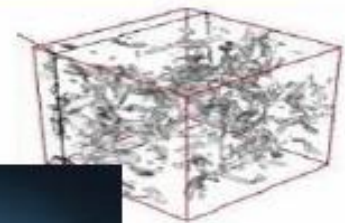
Évolutions de la science

Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
 - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
 - Newton's Laws, Maxwell's Equations...
3. Last few decades – **Computational Science**
 - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
 - Scientists overwhelmed with data sets from many different sources
 - Data captured by instruments
 - Data generated by simulations
 - Data generated by sensor networks
 - **eScience is the set of tools and technologies to support data federation and collaboration**
 - For analysis and data mining
 - For data visualization and exploration
 - For scholarly communication and dissemination



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



(With thanks to Jim Gray)



Evolution des pratiques de science

Science du 21^{ème} 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



5 programmes d'action de la SNRI

DU LABORATOIRE AU PATIENT

Associer recherche en laboratoire, recherche clinique et innovation privée pour le bénéfice des patients

BIOLOGIE DES SYSTÈMES ET APPLICATIONS

Soutenir l'émergence de concepts nouveaux pour la compréhension du vivant et développer les applications médicales et industrielles

SYSTÈME TERRE : OBSERVATION, PRÉVISION, ADAPTATION

Organiser l'acquisition et l'exploitation des données d'observation de la Terre pour anticiper les conséquences du changement climatique

BIG DATA

Un gisement exceptionnel de connaissances et de croissance

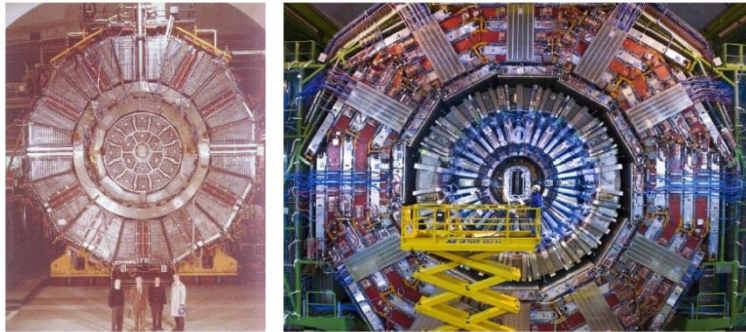
HOMME ET CULTURES

Appréhender les phénomènes humains dans leurs réalités individuelles et sociales



Il y a données et données....

In Big Communities In International Labs (CERN)



Past Century collaboration
~500 Scientists

Today collaboration
~4000 Scientists

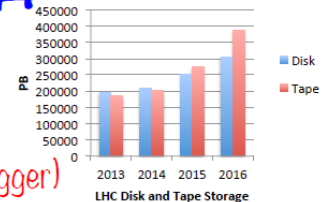
From all around the world

Données en HEP

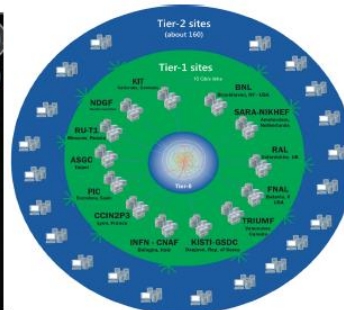
The (Big) DATA

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

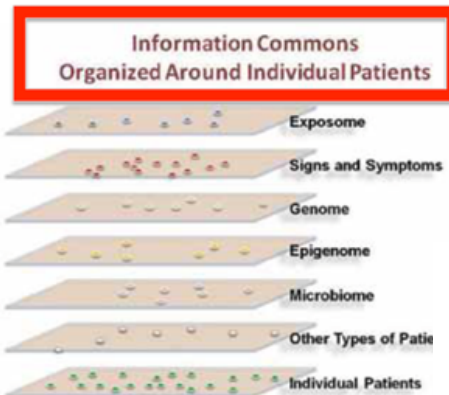
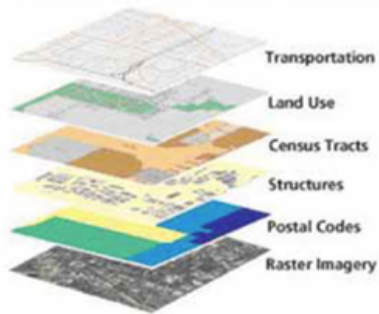
European Space Agency

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

Les données de santé

Intégration des données pour une Médecine translationnelle, prédictive et personnalisée

Google Maps: GIS layers
Organized by Geographical Positioning



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

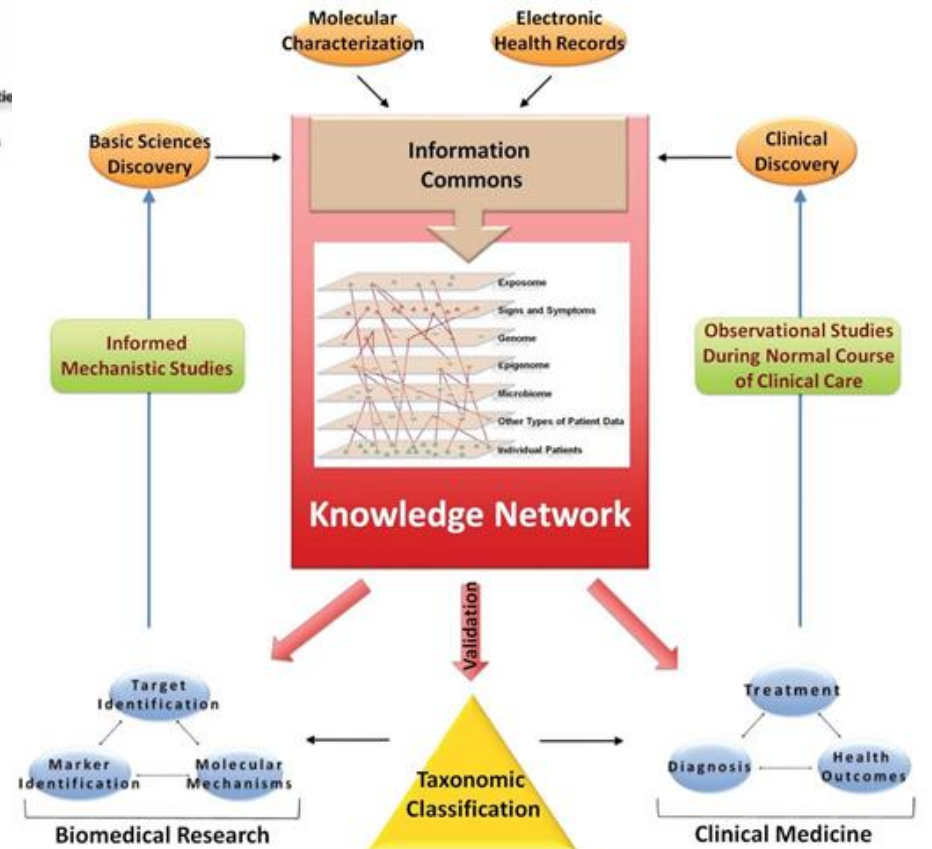
Marc CUGGIA (MD,PhD)

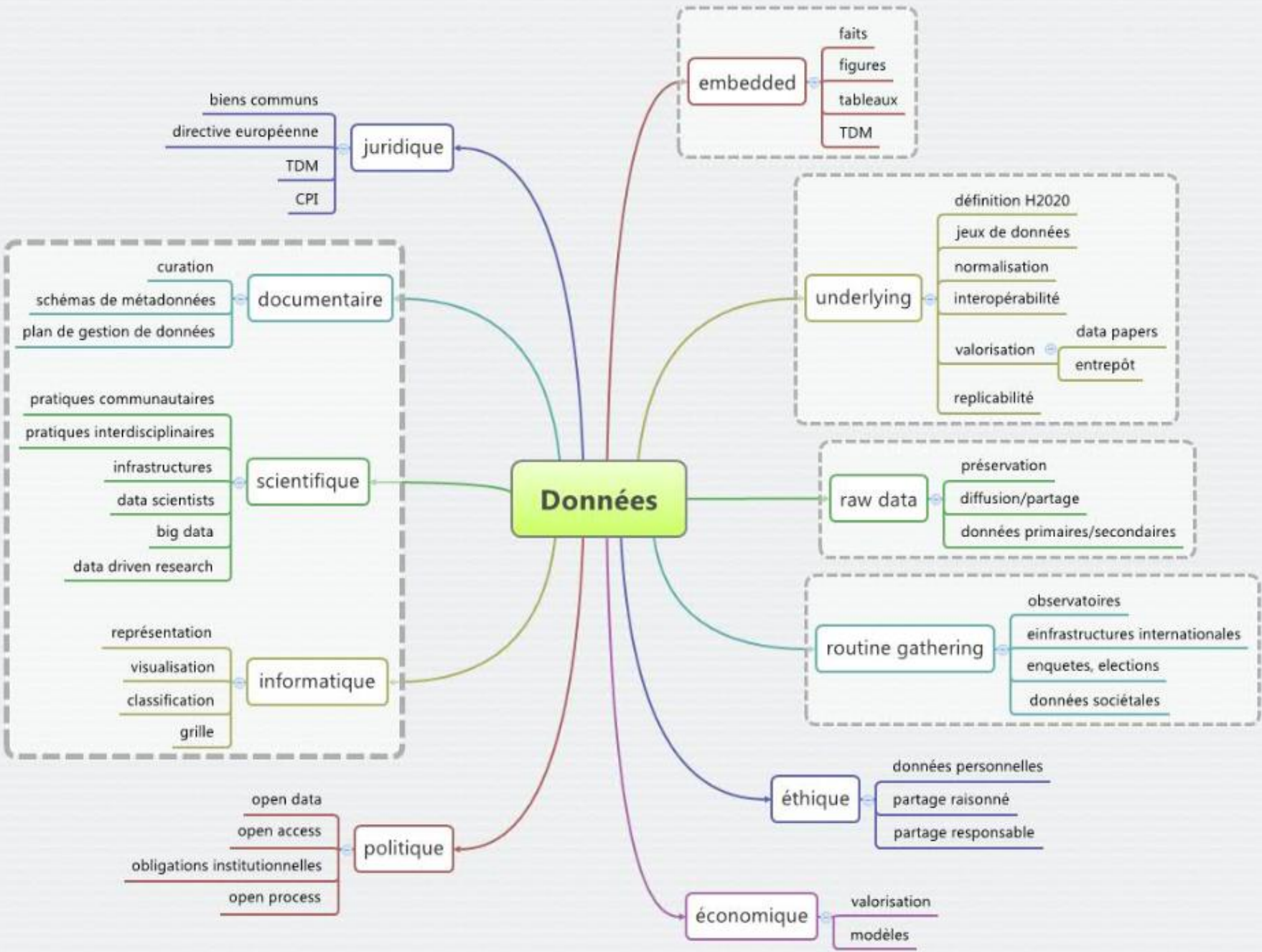
Health Big Data team (LTSI) -

Clinical Investigation Center (CHU Rennes)

INSERM – Medical School

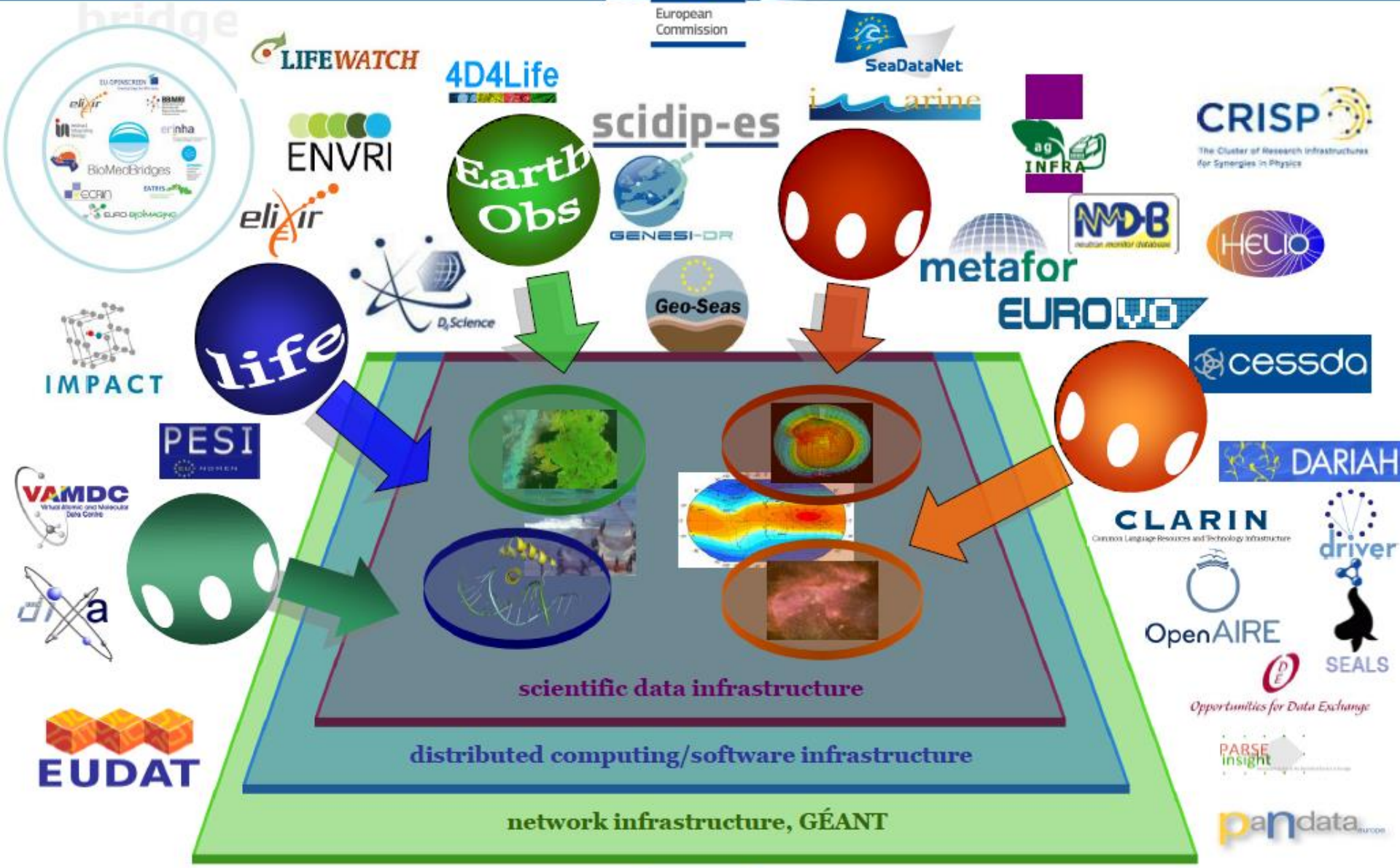
Université de Rennes 1 - BRITTANY



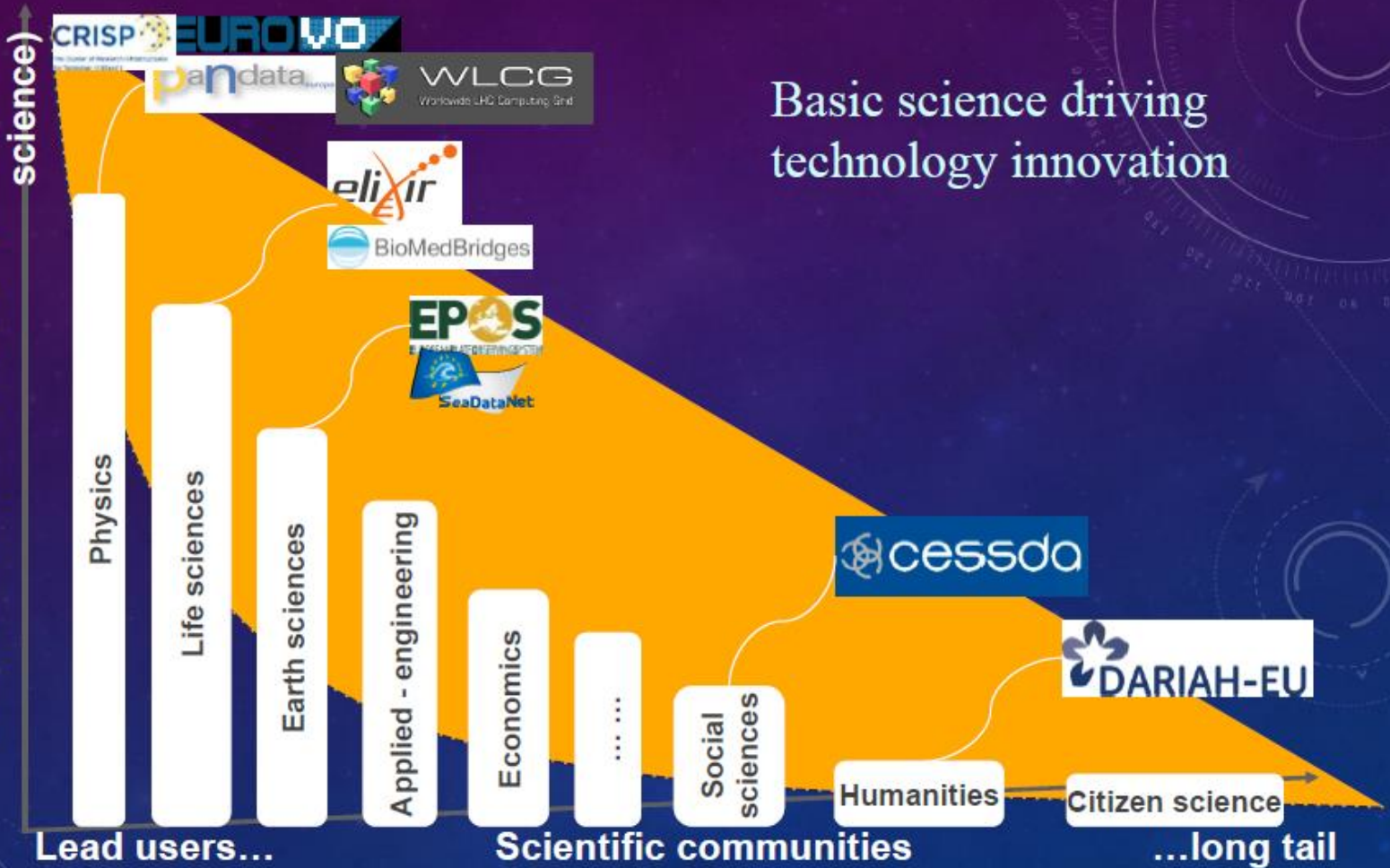


data infrastructure: bridging islands

Il est grand temps
de s'organiser...

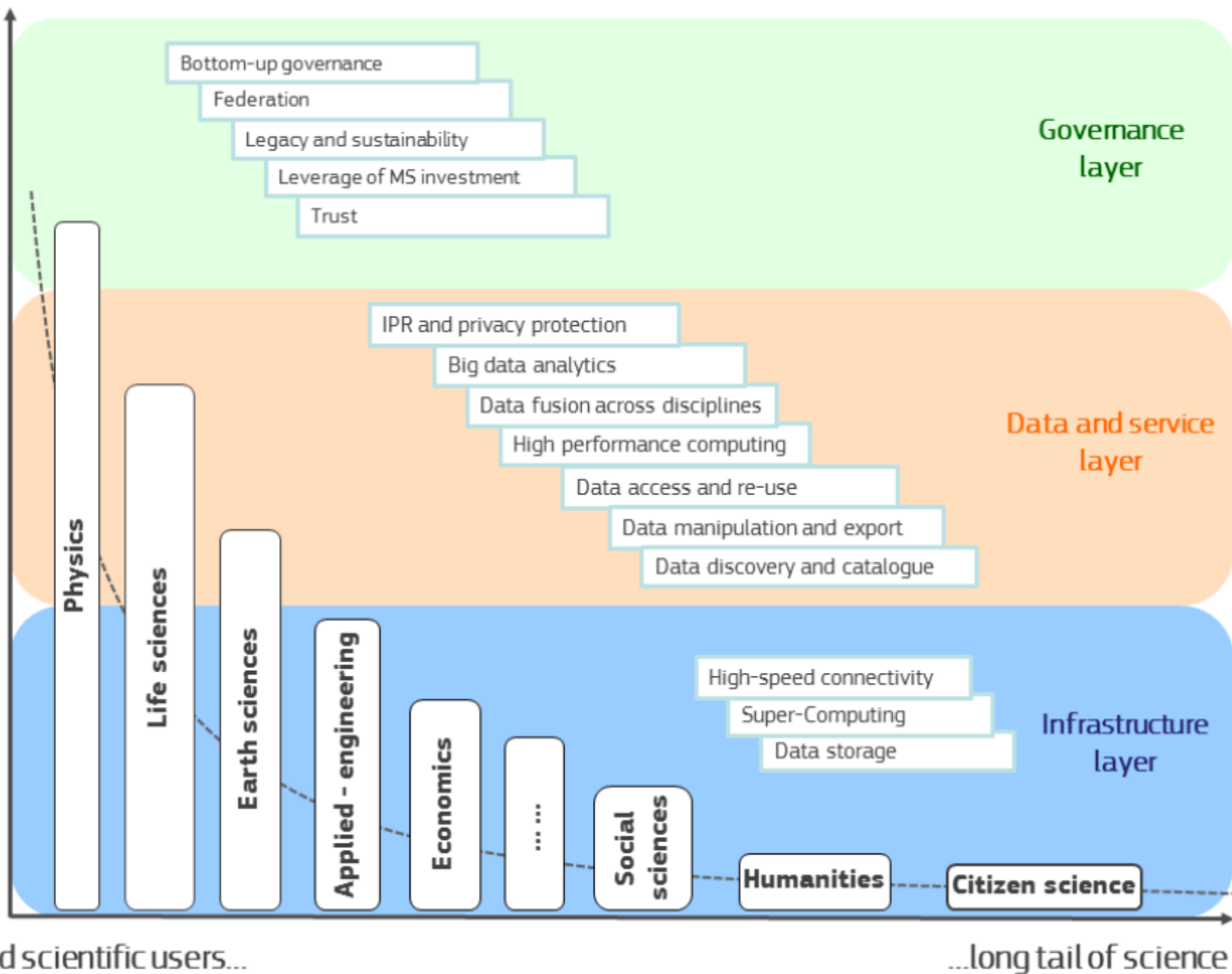


Scale of scientific activity (data-driven science)



Basic science driving technology innovation

Scale of scientific activity (data-driven science)





Vous avez dit Open Science ?

Un discours politique centré sur les technologies



Évolution des outils
et processus

New science:
New disciplines, new research topics

Intégration du citoyen

**More efficient
science:**
shared
resources,
dynamic
collaborations,
democratization
of research

**New research
methods,**
e-infrastructures,
big data

**Citizen
engagement in
research (citizen
science) and
scientific debate**

**ICT-enabled transformation of
science**

**Open research
collaborations**
(open science),
crowdsourcing

Open access
to research results
(publications &
data) and
processes

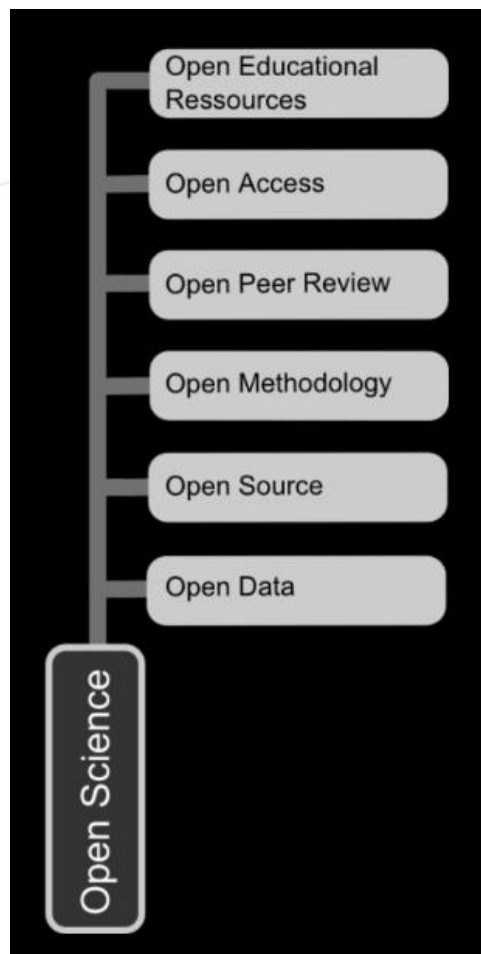
**Higher
impact
science:**
relevance,
accessibility
and impact
to society
and industry

Politique de coopération

Better science:
Transparent and replicable research

Changement de modèle
de diffusion

De multiples vues de l'Open science



Open Science aims at transforming science through ICT tools, networks and media, to make research more open, global, collaborative, creative and closer to society.

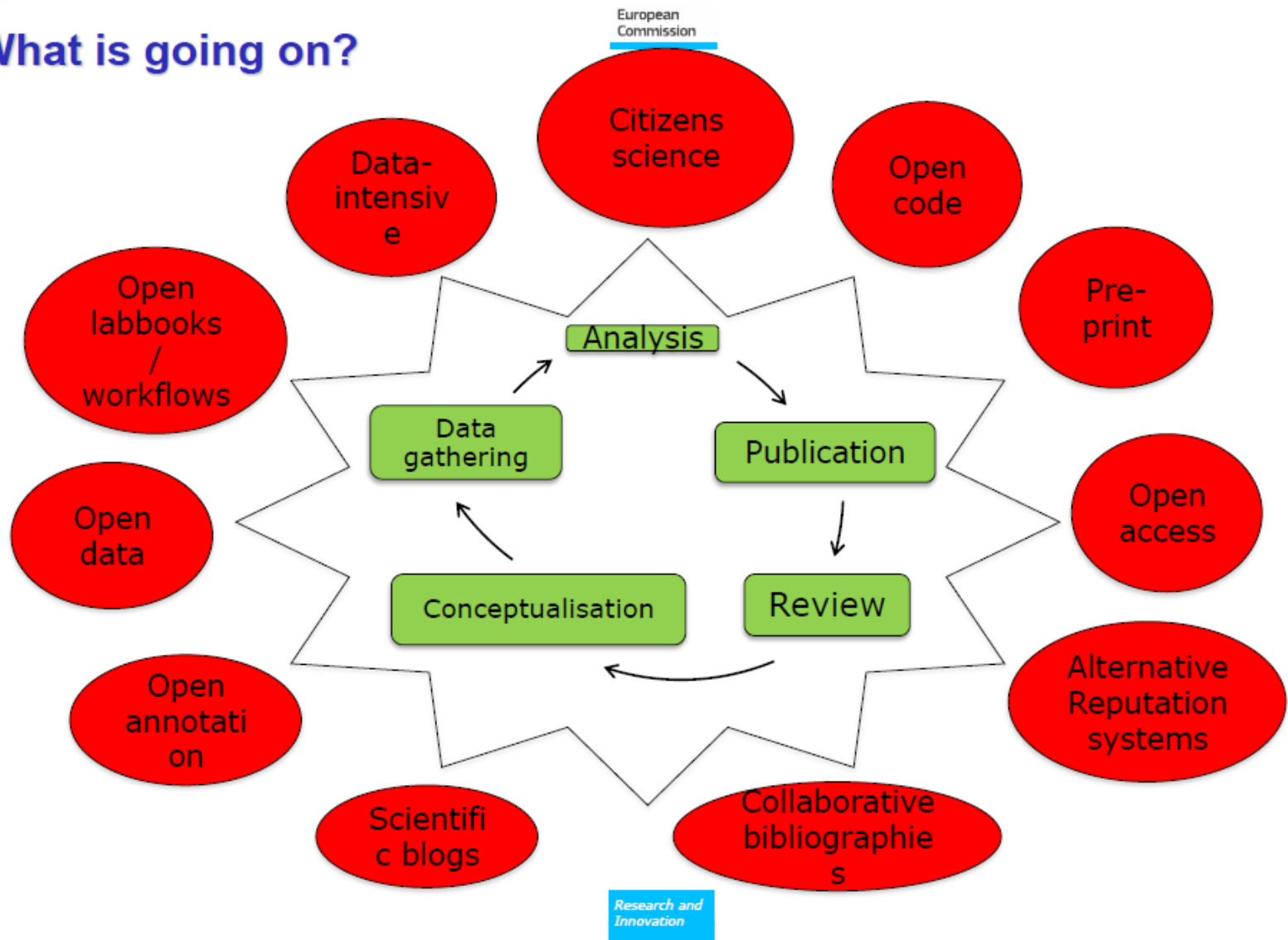
<https://ec.europa.eu/digital-agenda/en/open-science>

Open science is the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional

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

Science 2.0

2. What is going on?

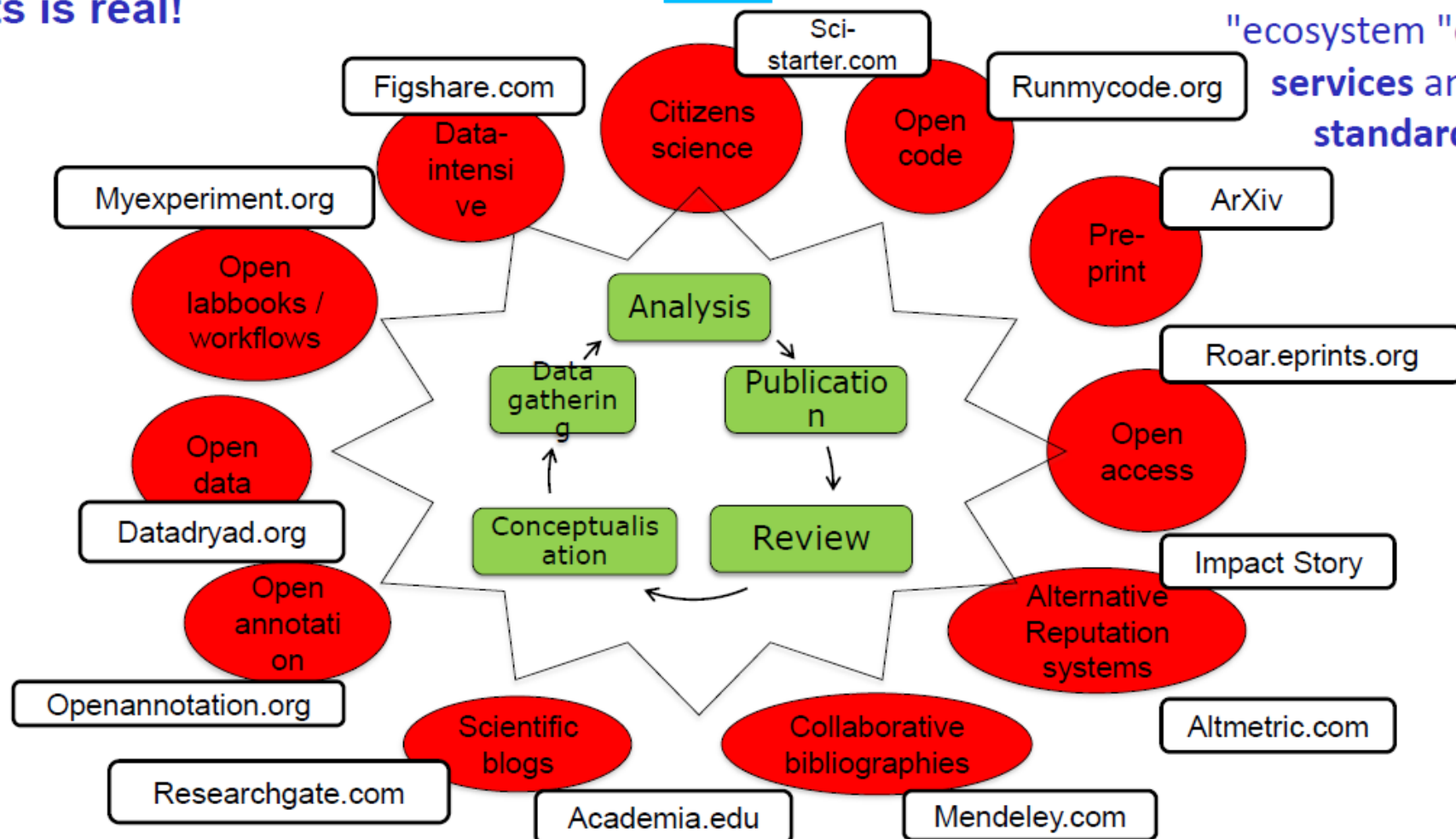


Its is real!

An emerging
"ecosystem" of
**services and
standards**

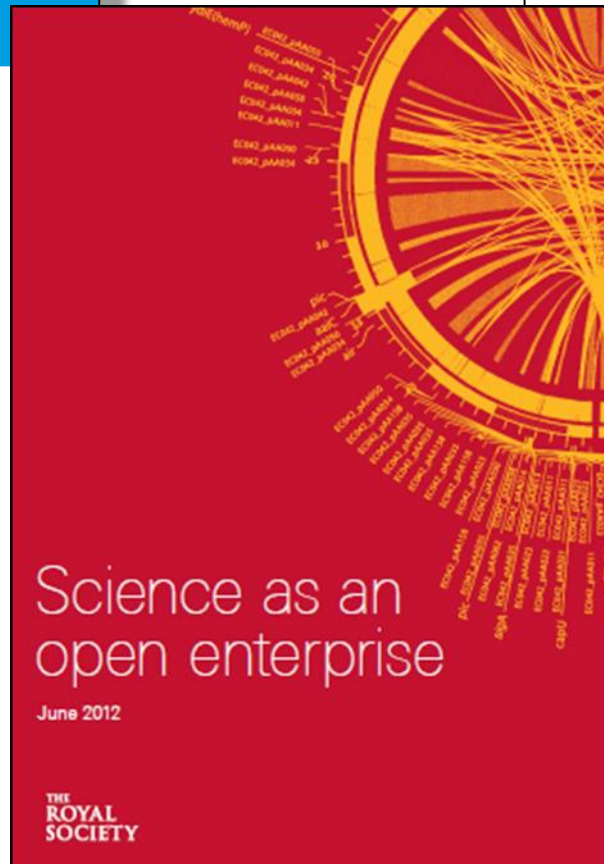
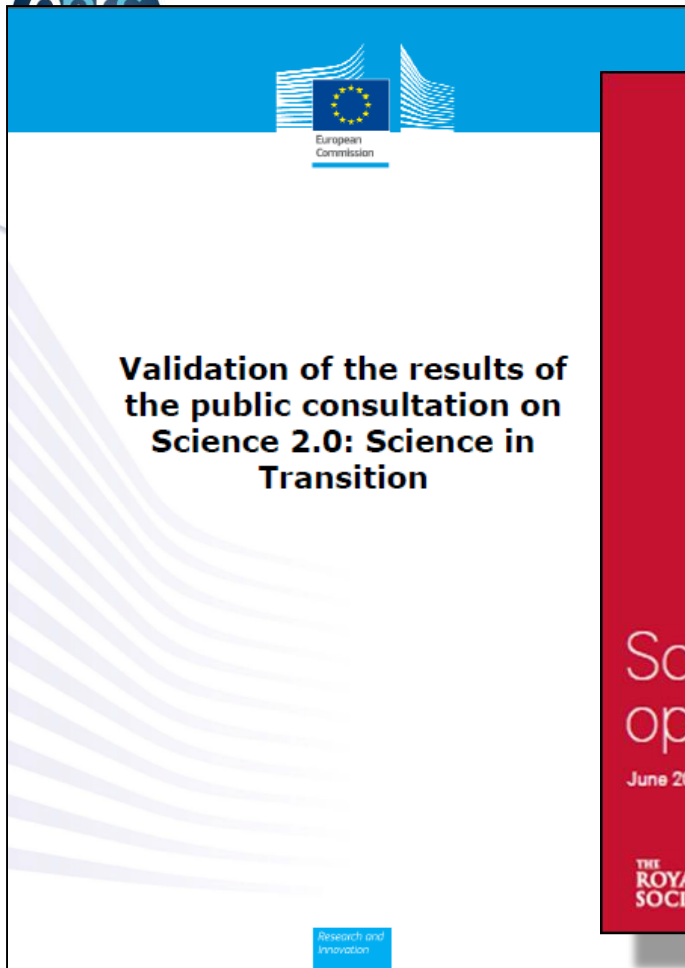
European
Commission

Research and
Innovation



Conseils de lecture...

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



<http://ec.europa.eu/digital-agenda/en/news/final-report-science-20-public-consultation>
<https://royalsociety.org/topics-policy/projects/science-publicenterprise/report/>



Developing a European Open Science Agenda: The Roadmap

Autumn 2015: *Stakeholder Discussion* on the European Open Science Agenda

Autumn 2015\Spring 2016: Concretization of Open Science actions under the *DSM strategy*

4-5 April 2016: Conference on Open Science during the Dutch Presidency

May 2016: Presentation of the European Open Science Agenda to the Competitiveness Council



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...

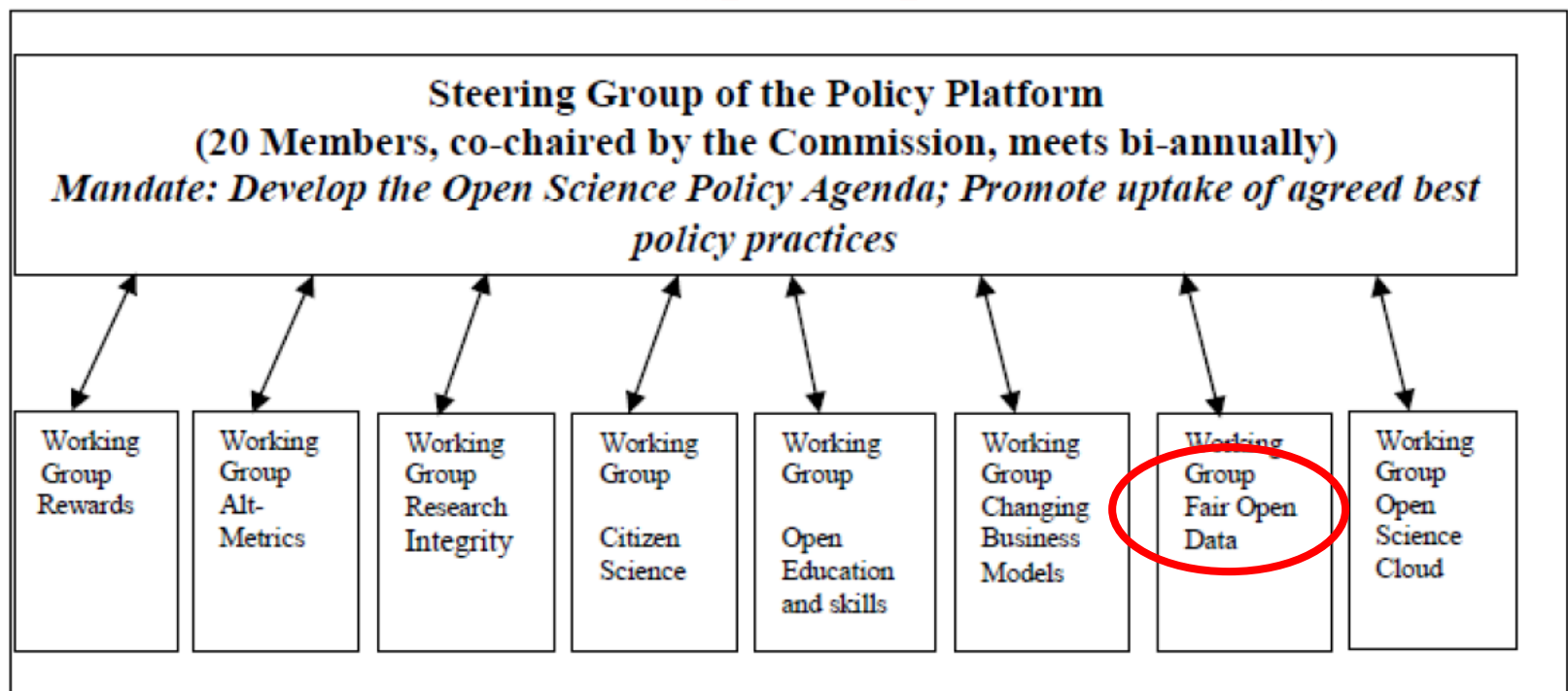
2016



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

Une infrastructure européenne de données estimée à 6,7 Md €



€2 BN IN OVERALL HORIZON 2020 FUNDING TO THE EUROPEAN CLOUD INITIATIVE, WITH ESTIMATED ADDITIONAL PUBLIC AND PRIVATE INVESTMENT OF €4.7 BN REQUIRED TO FURTHER DEVELOP THE EUROPEAN DATA INFRASTRUCTURE.

0.2bn
Widening the user
base to the public
and private sectors



1bn
EU-wide Quantum
technologies flagship

3.5 bn
Data Infrastructure

Conseil de lecture...

ERA n'est pas tombé du ciel bruxellois...



Key European stakeholders for research policies

ALLEA: the federation of All European Academies was founded in 1994 and currently brings together 56 academies in more than 40 countries from the Council of Europe region.

Business Europe: is an association founded in 1958 of 40 national business federations from 34 countries representing the interests of enterprises.

CESAER: the Conference of European Schools for Advanced Engineering Education and Research – is a non-profit international association of 50 European universities of technology and engineering schools/faculties from 24 countries, founded in 1990.

EARTO: the European Association of Research and Technology Organisations is an association of 91 members representing more than 350 research and technology organisations across the European Union and countries associated with the framework programme founded in 2000.

EIROForum: this forum of eight European Intergovernmental Research Organisations was established in November 2002 by the signature of a Charter.

ESF: the European Science Foundation was established in 1974 to act as a coordinating body for Europe's main research funding and research performing organisations. Its mandate was progressively modified by the creation of Science Europe and terminated in 2016.

EUA: the European University Association founded in 2001 represents 850 universities from 47 countries.

Eurodocs: the European Council of Doctoral Candidates and Junior Researchers is a federation of 35 national association representing early stage researchers founded in 2002.

Euroscience: is the non-profit grassroots researchers' association in Europe founded in 1997.

LERU: the League of European Research Universities is an association of 21 research-intensive universities from 10 countries founded in 2002.

Nordforsk: NordForsk was established in 2005 by the Nordic Council of Ministers and represents the actors of the research systems of the Nordic countries.

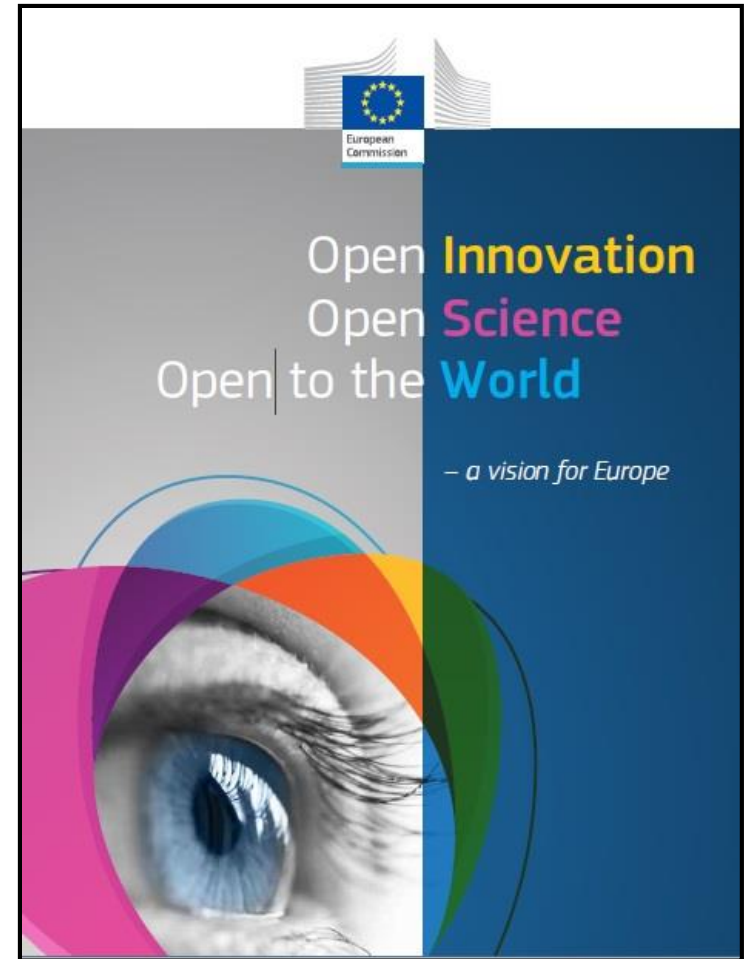
Science Europe: is an association of 47 European Research Funding Organisations and Research Performing Organisations from 27 countries, founded in October 2011. It included EUROHORCS, the European Heads of Research Councils association established in 1992 to represent the national research founding organisations (national research councils).

http://www.europarl.europa.eu/RegData/etudes/IDAN/2016/579097/EPRS_IDA%282016%29579097_EN.pdf



Carlos Moedas : une stratégie élargie

- Lien recherche >>> Innovation
- Partenariats Public Privé
- Ouverture vers la société
- Libre circulation des connaissances



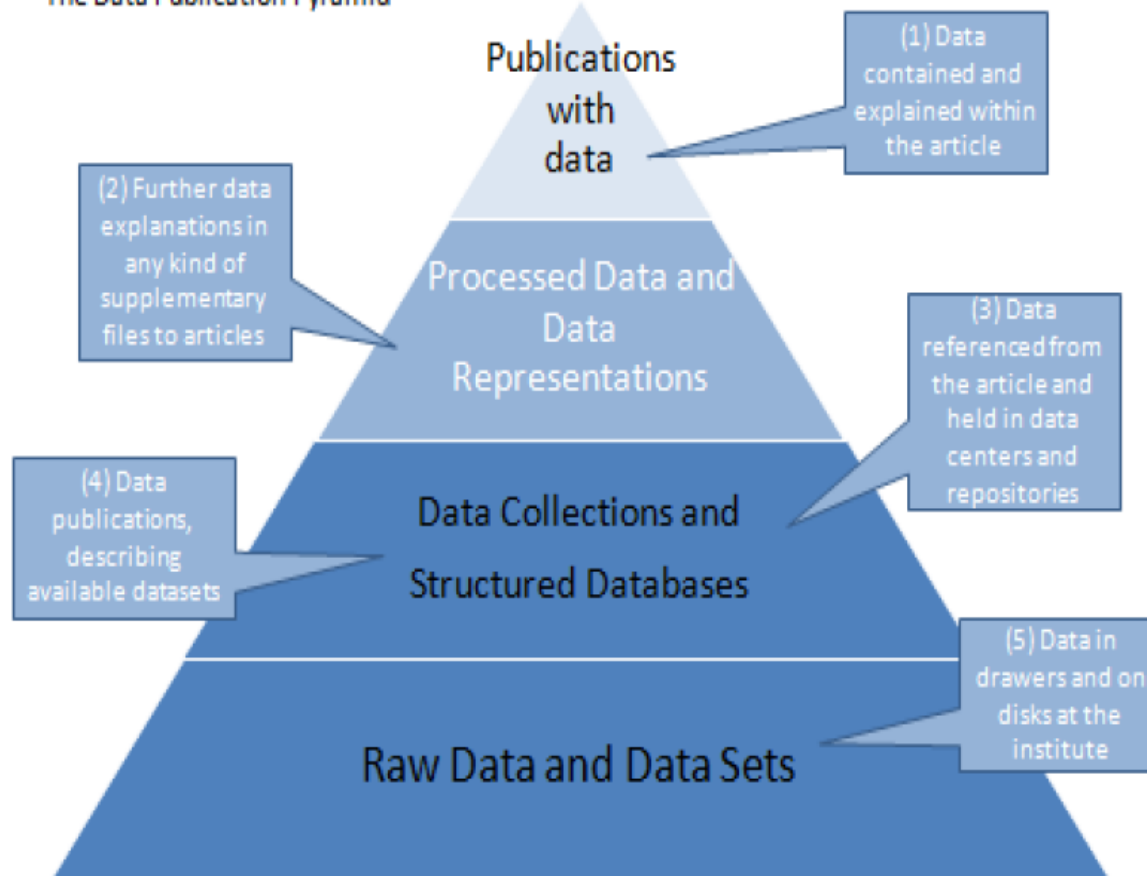


Classification des données

Où sont quelles données ?



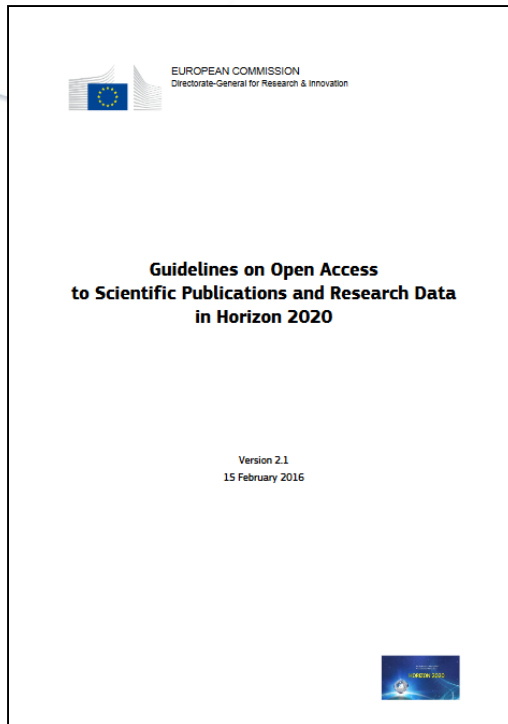
The Data Publication Pyramid



The ODE Data Publication Pyramid at
http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf



Horizon 2020 Open Access guidelines



Research data refers to information, in particular facts or numbers, collected to be examined and considered and **as a basis for reasoning, discussion, or calculation**. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form

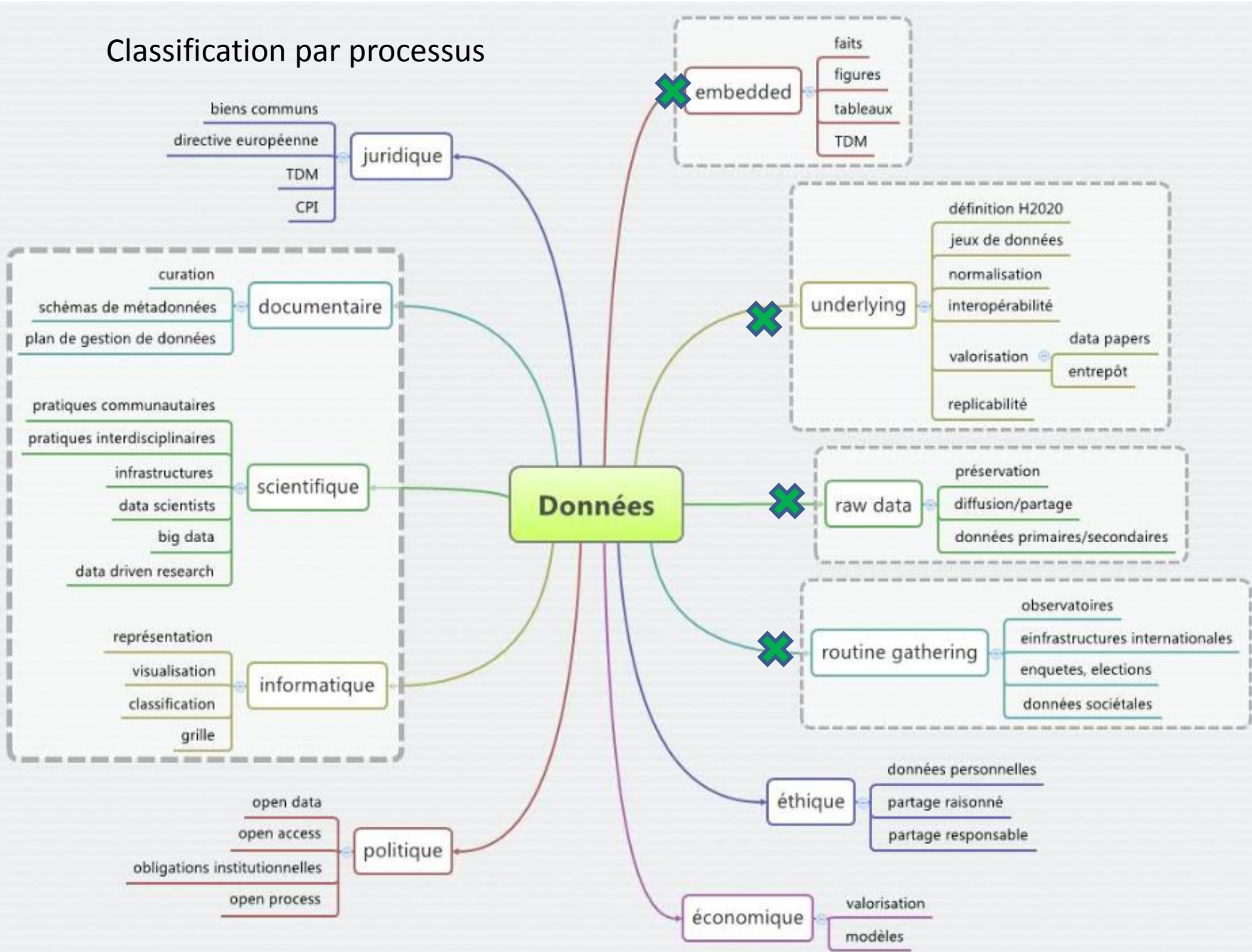
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Classification canonique des données de la recherche d'après le rapport de R. Gaillard, 2014, p18, citant la NSF et le RIN (Research Information Network)



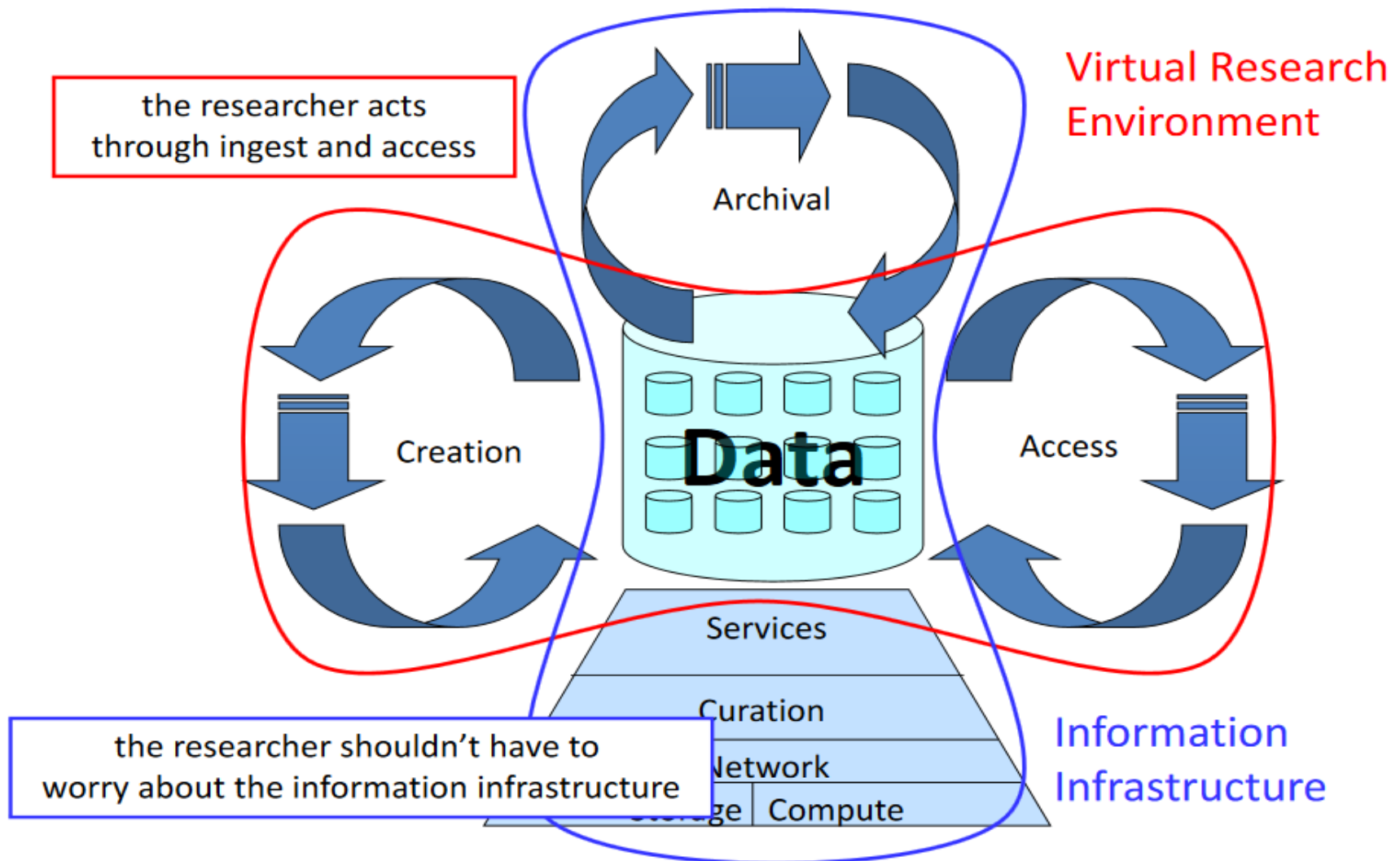
- **Données d'observation**, collectées à un instant, nécessitant un appareil descriptif conséquent (conditions, méthodologie, équipement, ...). Indissociables d'un contexte donné et uniques et impossibles à reproduire. A conserver de façon pérenne : neuroimagerie, concentration de phytoplanctons, cliché astronomique, données climatologiques, données d'enquête, séquence de gènes,
- **Données expérimentales**, obtenues à partir d'équipements suivant une méthodologie bien définie. Potentiellement reproductible, mais à des coûts parfois prohibitifs. La conservation dépend des investissements engagés dans leur production et de leur possible reproductibilité : chromatogrammes, cinétique chimique,
- **Données computationnelles ou de simulation**, issues de simulations à partir de modèles informatiques. Potentiellement reproductibles si le modèle informatique est correctement documenté : modèles de simulation sismique, modèles météorologiques, modèle économique, ...
- **Données dérivées ou compilées**, issues du traitement, de la combinaison ou de la réorganisation de données brutes, pour les rendre plus lisibles ou les présenter sous une forme canonique : imagerie IRM, fouille de texte, bases de données compilées

Classification par processus



La gestion des données ? Le truc dont les chercheurs ne veulent pas entendre parler...

Data centric view of research

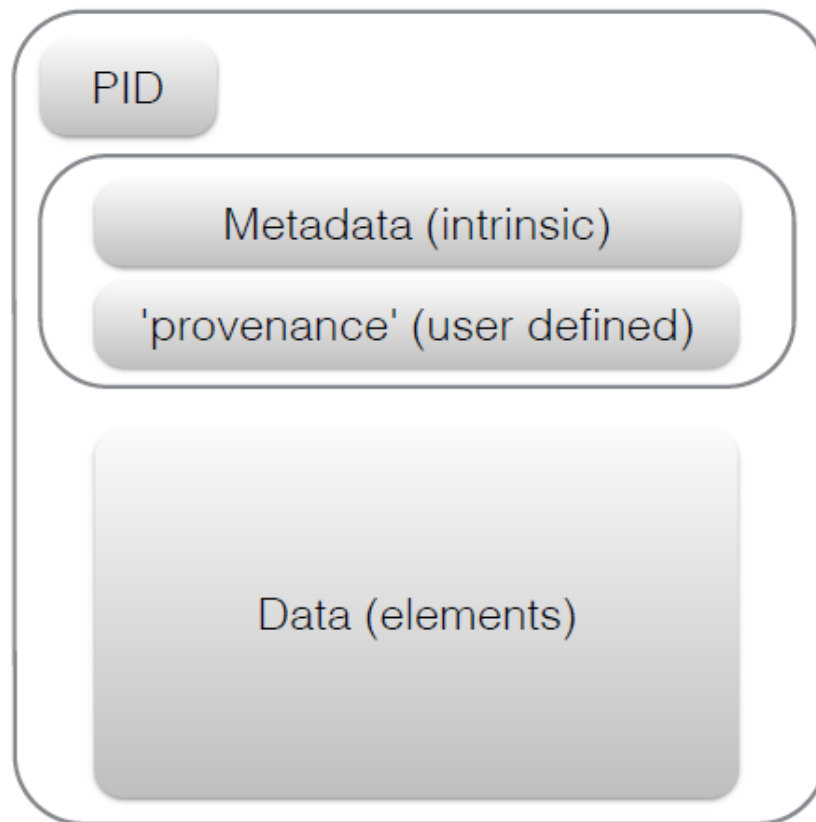




Qualité des données



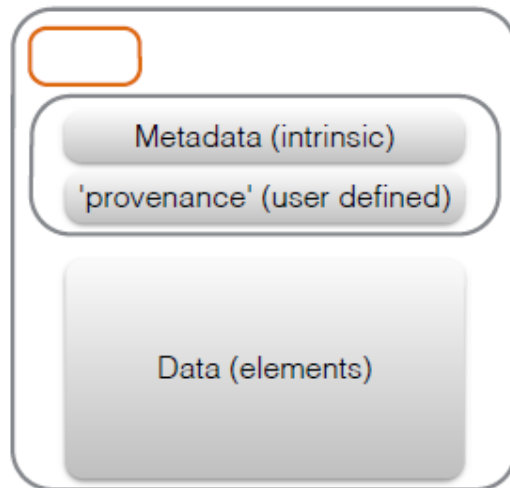
Corporation for National Research Initiatives



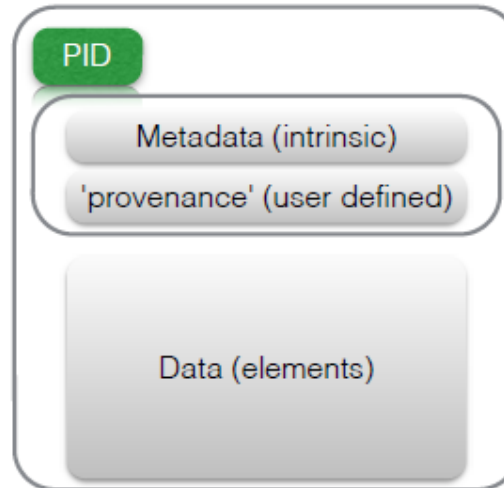
A simplified diagram of a Digital (data) Object irrespective of technological choices and naming

Data as increasingly FAIR Digital Objects

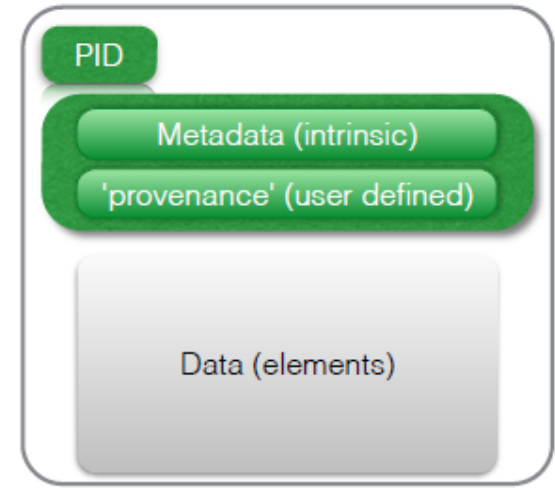
Totally UNFAIR



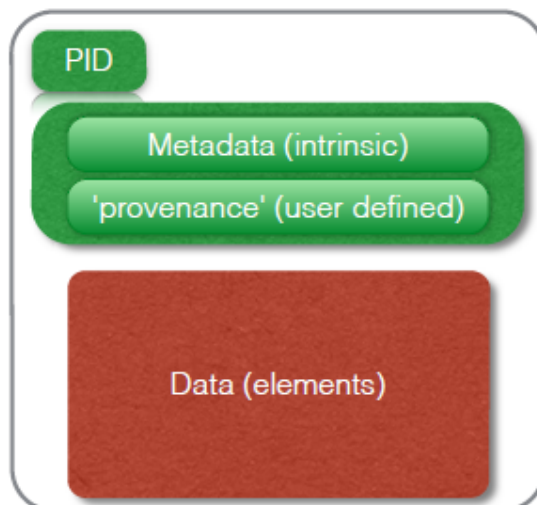
Findable
Usable for Humans



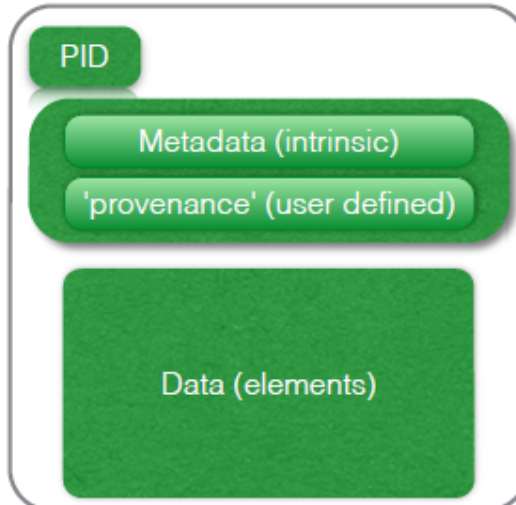
FAIR metadata



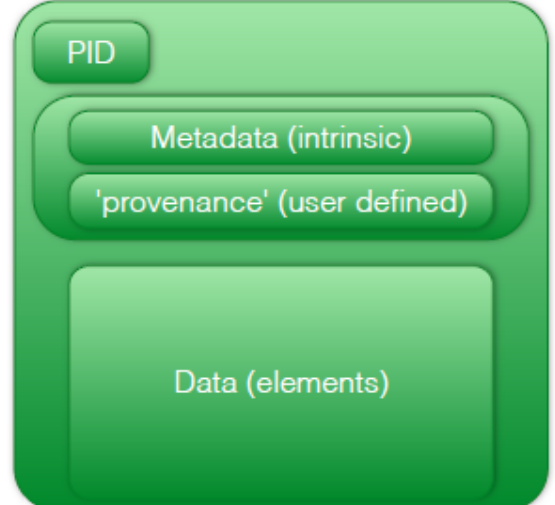
FAIR data-
restricted access



FAIR data-
Open Access



FAIR data-
Open Access/Functionally Linked



Où se niche la qualité des métadonnées ?



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

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**

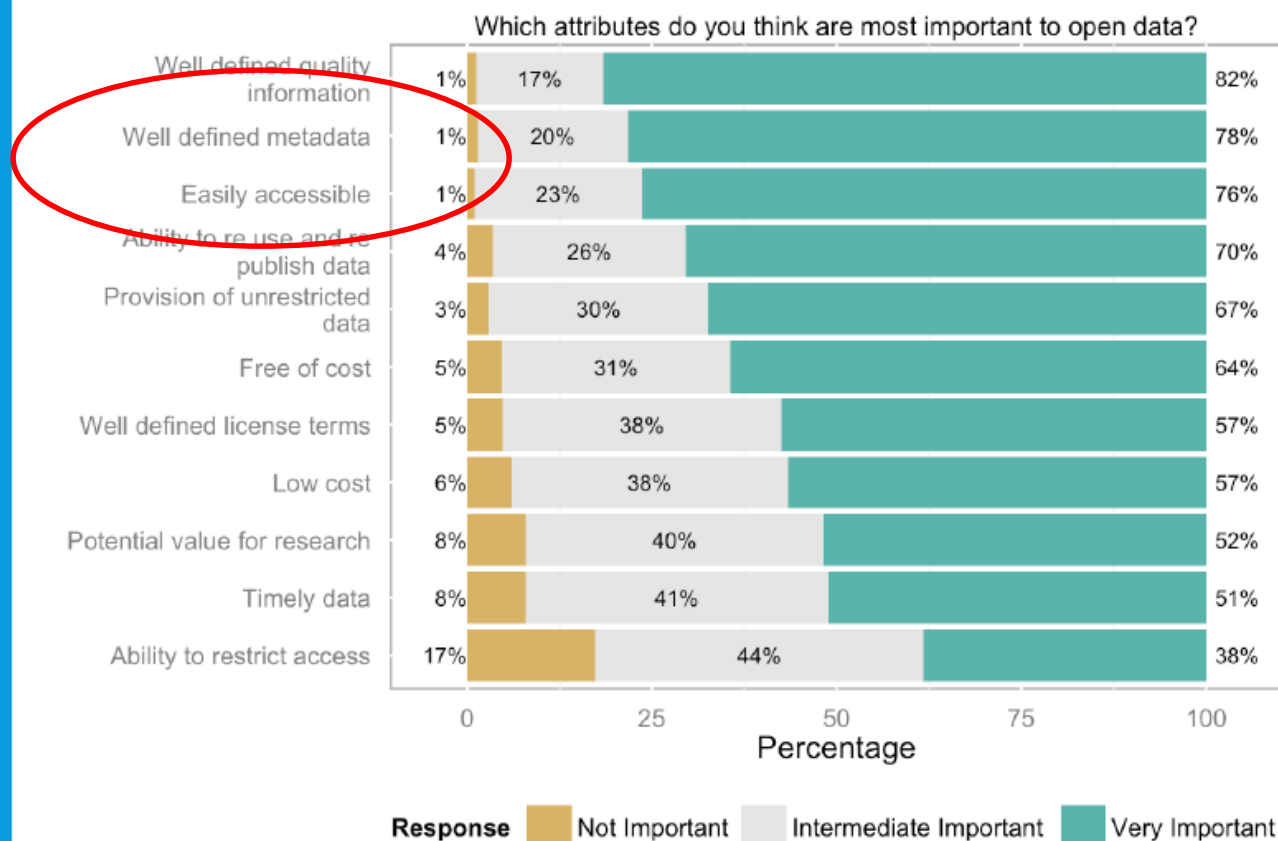
Moins habituel :

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



Données ouvertes : caractéristiques attendues

Qualité des données et des métadonnées





Culture du partage de la donnée



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



Vision:

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

Mission:

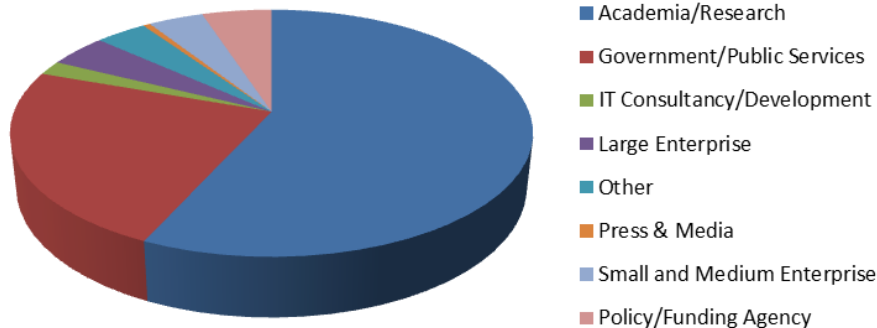
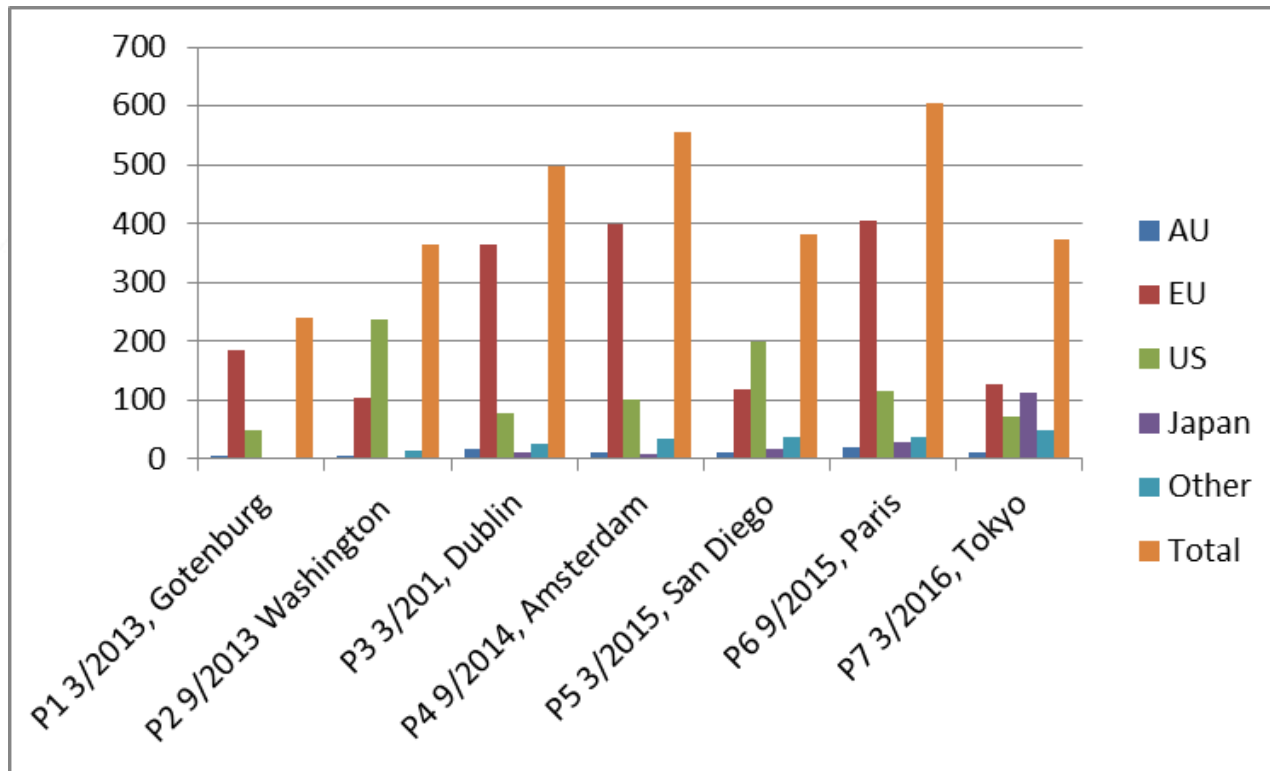
Building the social and technical **bridges**

Guiding Principles

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



Participation à RDA



- about 4000 registered members
- about 43 supporting organisations
- funders: EC, NSF/NIST, AU
 - more to come soon

RDA Europe Adoption Projects

EURO-ARGO
(Oceanography)



Dynamic Data Citation & the Argo data set
Adopting: Dynamic Data Citation

Wheat Interoperability
(Agriculture)



Creation of a Query interface for phenotyping data
Adopting: Wheat Data Interoperability (WDI)

Data Management
(Generic)



Integration of the RDA Metadata Standards Directory into DMPonline
Adopting: Metadata Standards Directory

CLARIN-Perseus
(Humanities)



Analysis of the OpenPhilology/Perseus and the CLARIN data repositories
Adopting: Data Foundation and Terminology

VAMDC
(Physics)



Implementation of a Query Store for the VAMDC infrastructure
Adopting: Data Citation

OpenAIRE
(Libraries)



Integration of the DLI Service into the OpenAire infrastructure
Adopting: Publishing Data services

Armenian
Infrastructure



Introduction of PIDs to the Armenian Life Sciences
Adopting: PID Information Types

Working Group examples

Data Infrastructure

1. Data Foundation and Terminology
2. PID Information Types
3. Data Type Registries
4. Practical Policy WG
5. Metadata Standards Directory
6. Brokering Governance
7. Dynamic Data Citation
8. Repository Platforms for Research Data*
9. Data Description Registry Interoperability

Data Publishing

1. RDA/WDS Publishing Data Bibliometrics
2. RDA/WDS Publishing Data Services

3. RDA/WDS Publishing Data Workflows

Education, Certification

1. RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World
2. Repository Audit and Certification DSA–WDS Partnership

Domain Specific WGs

1. The BioSharing Registry
2. Wheat Data Interoperability

PARTICIPEZ !

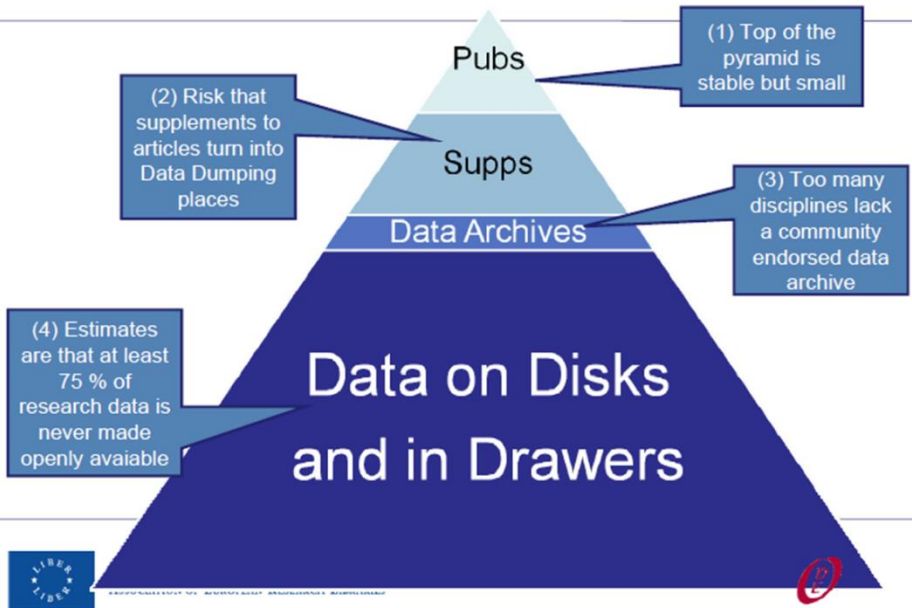
<https://rd-alliance.org/>



9

Et la suite ?

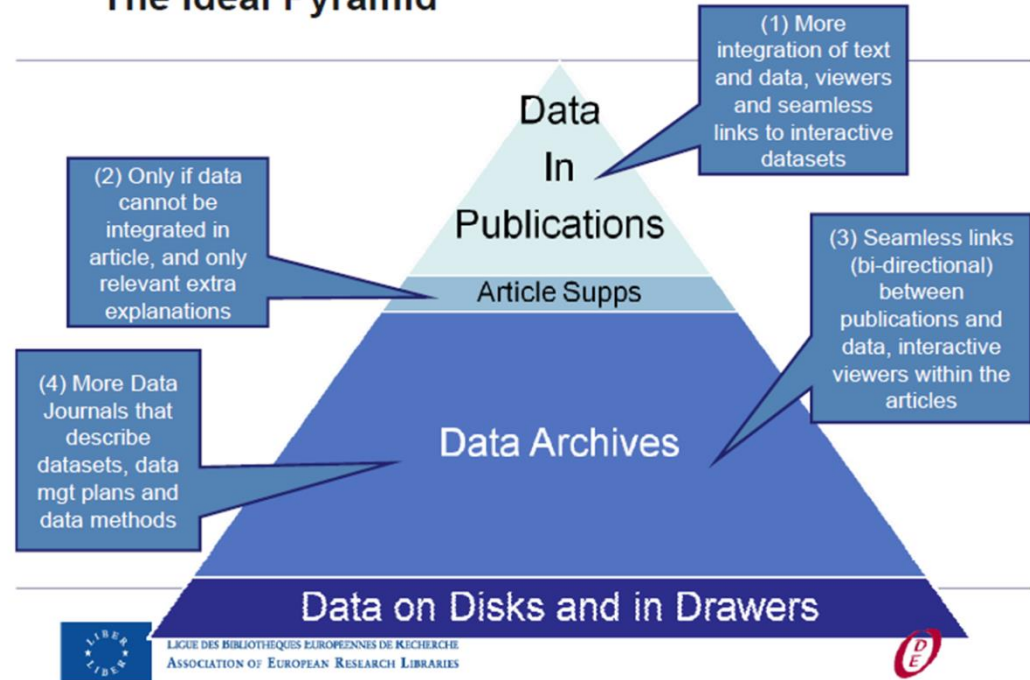
The Pyramid's likely short term reality:



Un (grand) pas à franchir...

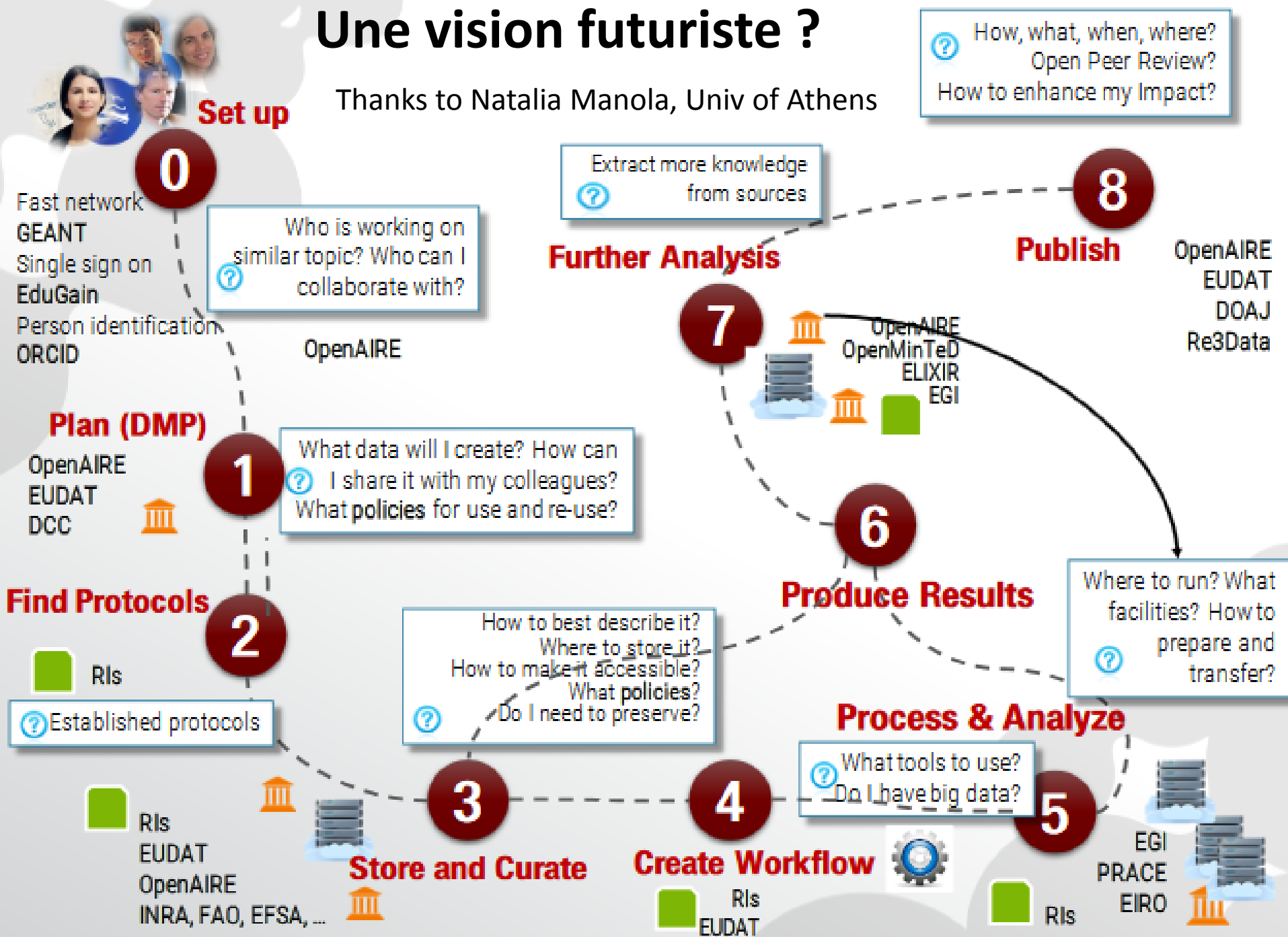
- Data journals
- Infrastructures disciplinaires
- Articles numériques : meilleure intégration texte/données

The Ideal Pyramid



Une vision futuriste ?

Thanks to Natalia Manola, Univ of Athens



Quelques points de discussion ?



- La plupart des enjeux 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 professionnels IST ont un rôle à jouer dans le développement des pratiques de confiance
- La valeur d'une production scientifique ne se limite pas à son utilisation première

Merci de votre attention

francis.andre@cnrs-dir.fr

Credits : Natalia Manola, Juan Bicarregui, Commission Européenne DG RDT, projet ODE, CNRI, Couperin, Monique Joly, CERN, SNRI, RDA, Jim Gray, Fran Berman, Marcello Maggi, Paul Tiesinga, Marc Cuggia,...