Open access to research data in a European policy context

Daniel Spichtinger, Senior Policy Officer, Unit A.6. (Science policy, foresight & data) DG Research & Innovation
daniel.spichtinger@ec.europa.eu

Jarkko Siren, Programme Officer, Unit C.1. (e-Infrastructure) DG Communications Networks, Content and Technology
jarkko.siren@ec.europa.eu

Open Access

- Online access at no charge to the *user*
  - to peer-reviewed scientific publications
  - to research data

Open Science

- A systemic change in the modus operandi of science and research
- Affecting the whole research cycle and its stakeholders

Commission objective

optimise the impact of publicly-funded scientific research

- At Member State level
- At European level (FP7 & Horizon 2020)

One way to get there: open access

Expected benefits:

- Better and more efficient science *Science 2.0*
- Economic growth *Innovation Union*
- Broader, faster, more transparent and equal access for the benefit of researchers, industry and citizens *Responsible Research and Innovation*

… in the European Research Area and beyond

A new Commission (2014-19)

Andrus Ansip, Vice-President, Digital Single Market

Günther Oettinger, Commissioner for Digital Economy and Society

Carlos Moedas, Commissioner for Research, Science and Innovation

Commissioner view

"Open Science, of which Open Access is an important part, will be vital to ensuring European progress and prosperity in the future"

(Speech at NETHER, January 26, 2015)

Open access policies across the EU

Preliminary findings from:

(i) National Points of Reference reporting template of 13 EU MS & 1 Associated Country

(ii) Results of 2014 ERA Progress Report

General findings

- Mostly soft measures rather than legislation: exceptions in more proactive and advanced OA policies
- OA to publications still much more developed than OA to data. Progress as to infrastructures for data (repositories), but openness still quite complex an issue and not addressed in many countries (for data)
- Bigger countries and countries with better budget capabilities tend to have more comprehensive OA policies and OA enabling infrastructures, as well as tend to lead or participate more actively in OA networking initiatives
- Nevertheless, smaller or less decentralised countries have the advantage of easier coordination, easier development of synergies

OA to research data in the MS

Some progress on infrastructure for data collection, archiving, access and re-use (But no overarching solution - challenges: different types of data, privacy, IP protection, funding for long-term preservation, standardisation, interoperability)

Some MS very proactively driving forward, contributing or participating in European and global organisations such as Knowledge Exchange, Science Europe, RDA, GRC, etc.

Bottom-up movement: Stakeholders such as research institutes and funders, universities and libraries, usually very actively driving OA forward

Top-down movement: although a need for coordination at EU level is perceived, many MS still need to formulate clear and comprehensive OA policies, strategies and laws

Open access in Horizon 2020

From FP7 to H2020: OA to publications from pilot to underlying principle

FP7

- **Green** open access pilot in 7 areas of FP7 with 'best effort' stipulation
- Allowed embargoes: 6/12 months
- **Gold** open access costs eligible for reimbursement as part of the project budget while the project runs

Horizon 2020

- **Obligation** to provide OA, either through the **Green** or **Gold** way in **all areas**
- Allowed embargoes: 6/12 months
- **Gold** open access costs eligible for reimbursement as part of the project budget while the project runs & **post-grant support being piloted**
- Aim to deposit at the same time the research data needed to validate the results **("underlying data")**

Pilot on Open Research Data in H2020

Three key questions:

Which thematic areas should be covered?

What kind of data should be covered?

What about data management?

Pilot on Open Research Data: Scope

Areas of the 2014-2015 Work Programme participating in the Open Research Data Pilot are:

- Future and Emerging Technologies
- Research infrastructures – part e-Infrastructures
- Leadership in enabling and industrial technologies – Information and Communication Technologies
- Societal Challenge: Secure, Clean and Efficient Energy – part Smart cities and communities
- Societal Challenge: Climate Action, Environment, Resource Efficiency and Raw materials – except raw materials
- Societal Challenge: Europe in a changing world – inclusive, innovative and reflective Societies
- Science with and for Society

Projects in other areas can participate on a voluntary basis.

Pilot on Open Research Data: Opt-out

Projects may opt out of the Pilot on Open Research Data in Horizon 2020 in a series of cases (submission stage):

- If the project will not generate / collect any data
- Conflict with obligation to protect results
- Conflict with confidentiality obligations
- Conflict with security obligations
- Conflict with rules on protection of personal data
- If the achievement of the action's main objective would be jeopardised by making specific parts of the research data openly accessible (to be explained in data management plan)

Pilot on Open Research Data: requirements

Types of data concerned:

- Data needed to validate the results presented in scientific publications ("underlying data")
- Other data as specified in data management plan (=up to projects)

Beneficiaries participating in the Pilot will:

- Deposit this data in a research data repository of their choice
- Take measures to make it possible to access, mine, exploit, reproduce and disseminate free of charge
- Provide information about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (where possible, provide the tools and instruments themselves)

EC: Support & monitoring (Annotated MGA, Specific guidance etc…)

Data management in Horizon 2020

- Data Management Plans (DMPs) mandatory for all projects participating in the Pilot, optional for others
  - DMPs are NOT part of the proposal evaluation, they need to be generated within the first six months of the project and updated as needed
- DMP questions:
  - What data will be collected / generated?
  - What standards will be used / how will metadata be generated?
  - What data will be exploited? What data will be shared / made open?
  - How will data be curated and preserved?

ORD Pilot: initial take-up in first calls of H2020

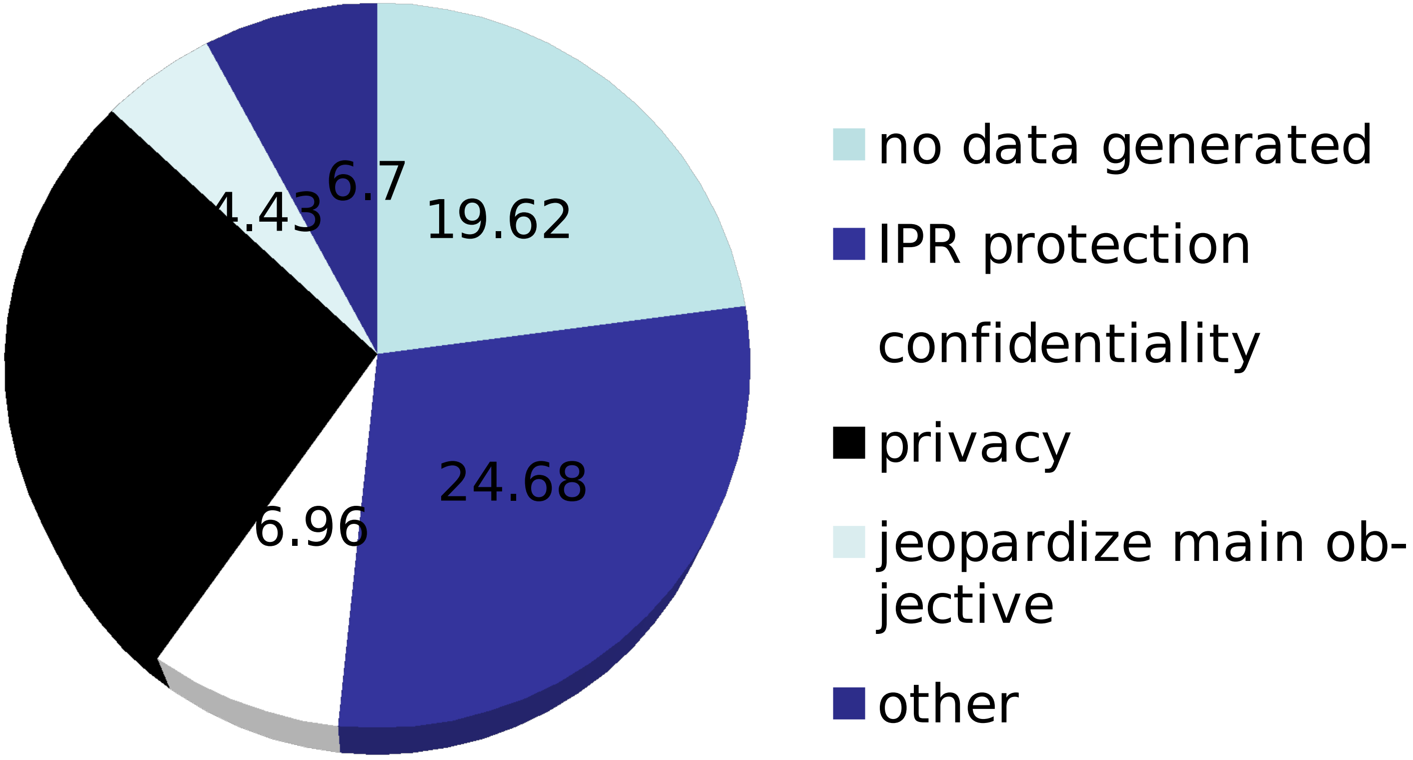
- Preliminary!
- Basis: 3054 Horizon 2020 proposals

Calls in core-areas: **opt out 24.2%** (442 of 1824 proposals) – range from 9,1-29,1%

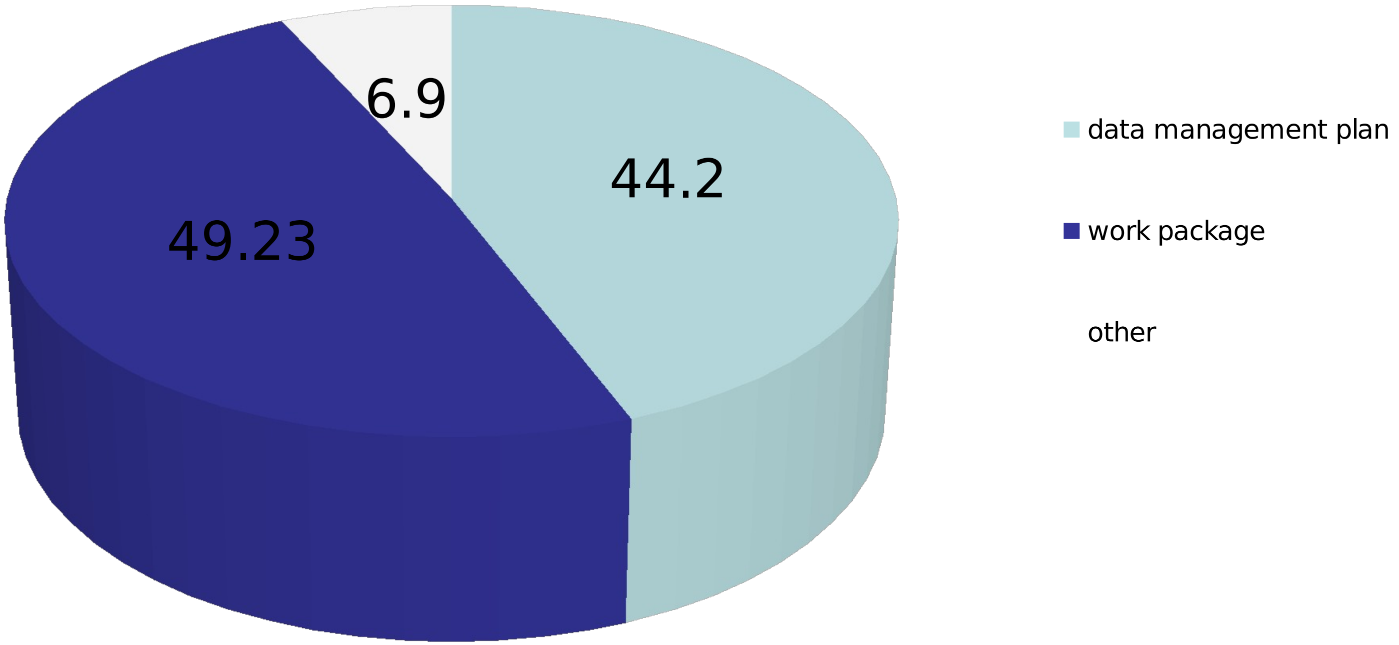
Other areas: voluntary **opt in 27.2%** (334 of 1230 proposals) – range from 9 to 50%

- Conclusion: 'early days' for the open access to research data pilot, but initial data on uptake in the proposals for the first calls of Horizon 2020 are encouraging. Initial areas well chosen (drop outs below 30%, similar range), comprehensive follow up needed

ORD Pilot: opt-out reasons among proposals



ORD Pilot: approach to data management among proposals



Further information: Science Metrix Study

State-of-art analysis of OA strategies to peer-review publications

State-of-art analysis of OA strategies to scientific data

Comparative analysis of the strengths and weaknesses of existing open access strategies

http://science-metrix.com/en/publications/reports

Ongoing coordination and support actions (FP7 funded)

PASTEUR4OA (Open Access Policy Alignment Strategies for European Union Research) Started 2014

FOSTER (Foster Open Science Training for European Research) Started 2014

RECODE - (Policy Recommendations for Open Access to Research Data in Europe) – 2013, finishing

OpenAIRE/OpenAIRE+: supporting the implementation of Open Access in Europe (publications and data)

Infrastructure projects(with OA components), e.g. GEO/GEOSS, ELIXIR…

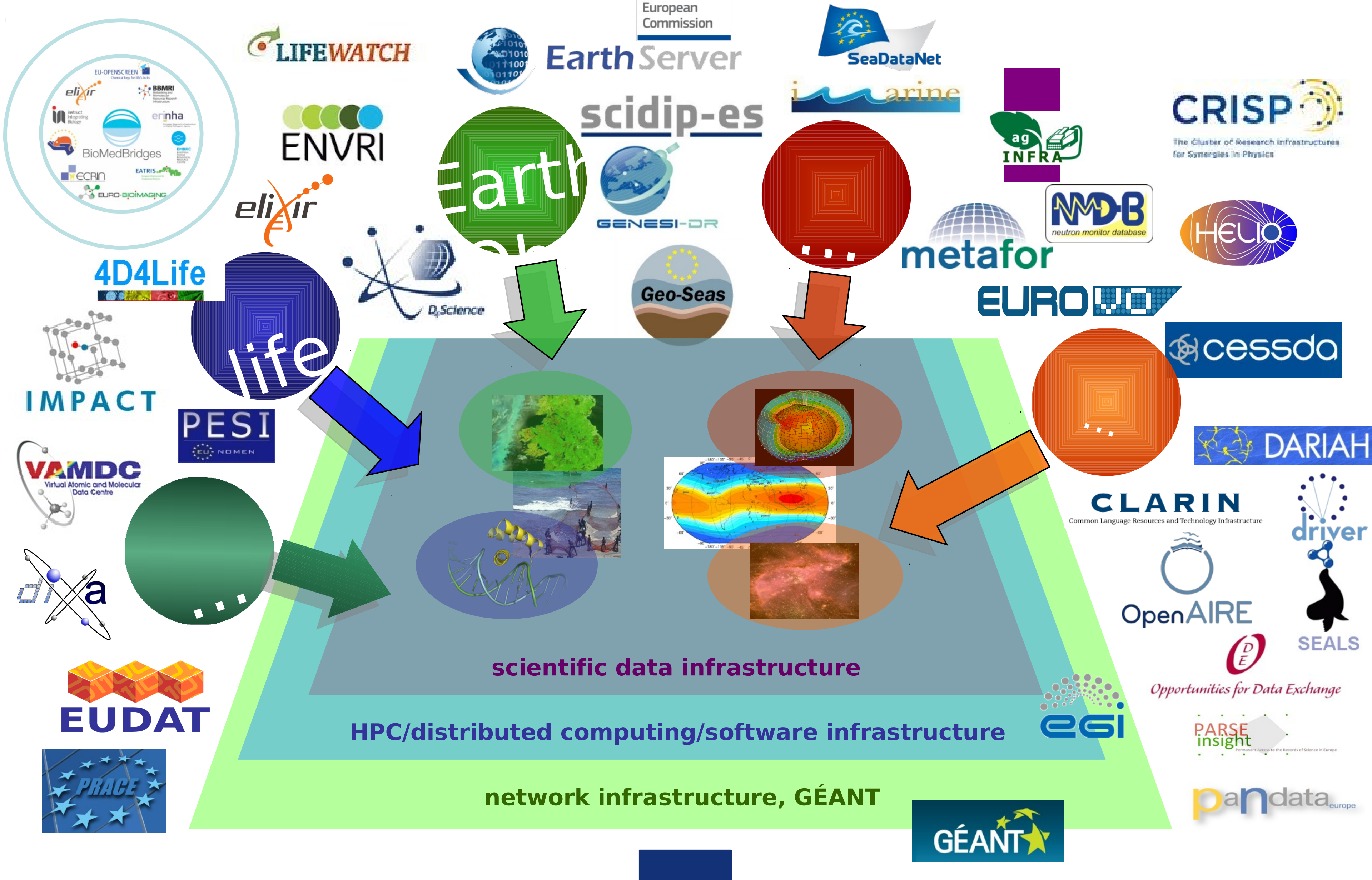
Open research data: challenges

- **Uptake of Open Research Data Pilot**: in signed grant agreements (versus proposals)
- **Structure and coverage of the pilot**: to remain the same, at least until mid-term review; small incremental changes possible
- **DMP implementation:** investigating best-practice; tools to be developed (2015)
- **Top-notch monitoring of OA policies** is crucial for further policy development

E-Science

- **Large scale collaborations becoming the norm**
    - *often global*
    - *virtual research communities*
    - *virtual access to rare/remote resources*
- **Data-intensive science and innovation**
    - *Use and manage exponentially growing, heterogeneous sets of data*
- **Experimentation in silico, simulation**
    - *Use of HPC, grid and cloud computing*

e-infrastructure building bridges



scientific data infrastructure

HPC/distributed computing/software infrastructure

network infrastructure, GÉANT

Issues to be addressed (data infrastructure)

The EC in coordination with EU Member States is looking after research data as an infrastructure

As a valuable and a strategic resource, research data opens at least three key issues to be addressed(*):

- **How data can be networked**
- **How to envision and set up data governance on a global scale**
- **How the EU can play a leading role in helping start and steer this global trend**

(*) Fred Friend, Jean-Claude Guédon Herbert van Sompel
"Beyond Sharing and Re-using: Toward Global Data Networking"

Research Data Alliance (RDA)

- Global data practitioner forum
- Activity takes place in work groups (e.g. "The BioSharing Registry: connecting data policies, standards & databases in life sciences")

Research Data Alliance
Research Data Sharing

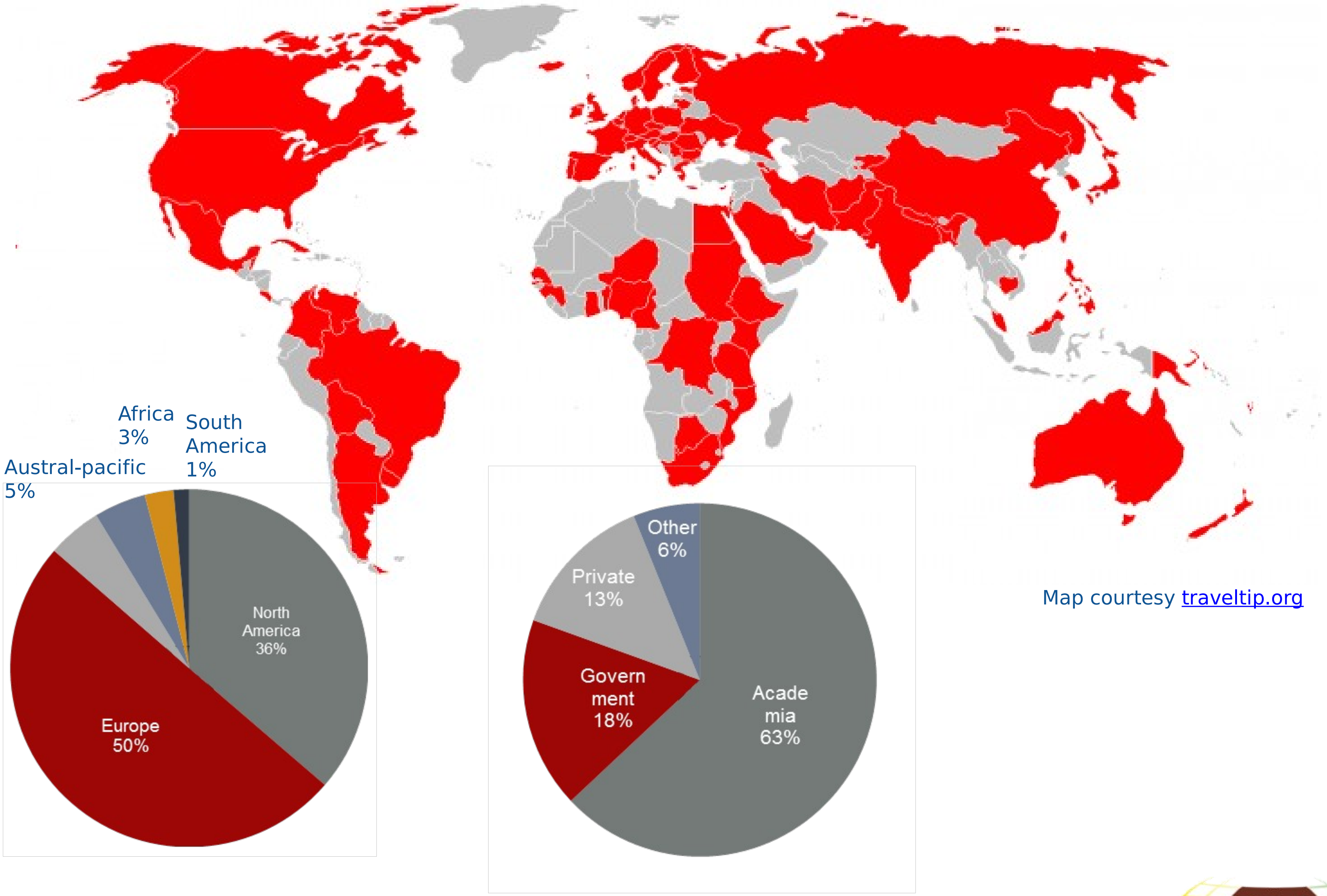
- **RDA community focuses on building social, organizational and technical infrastructure to**
  - **reduce barriers to data sharing and exchange**
  - **accelerate the development of coordinated global data infrastructure**

Plenary 2
Washington, DC

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



Map courtesy traveltip.org

Distribution of 2,353 Individual RDA Members in 96 Countries

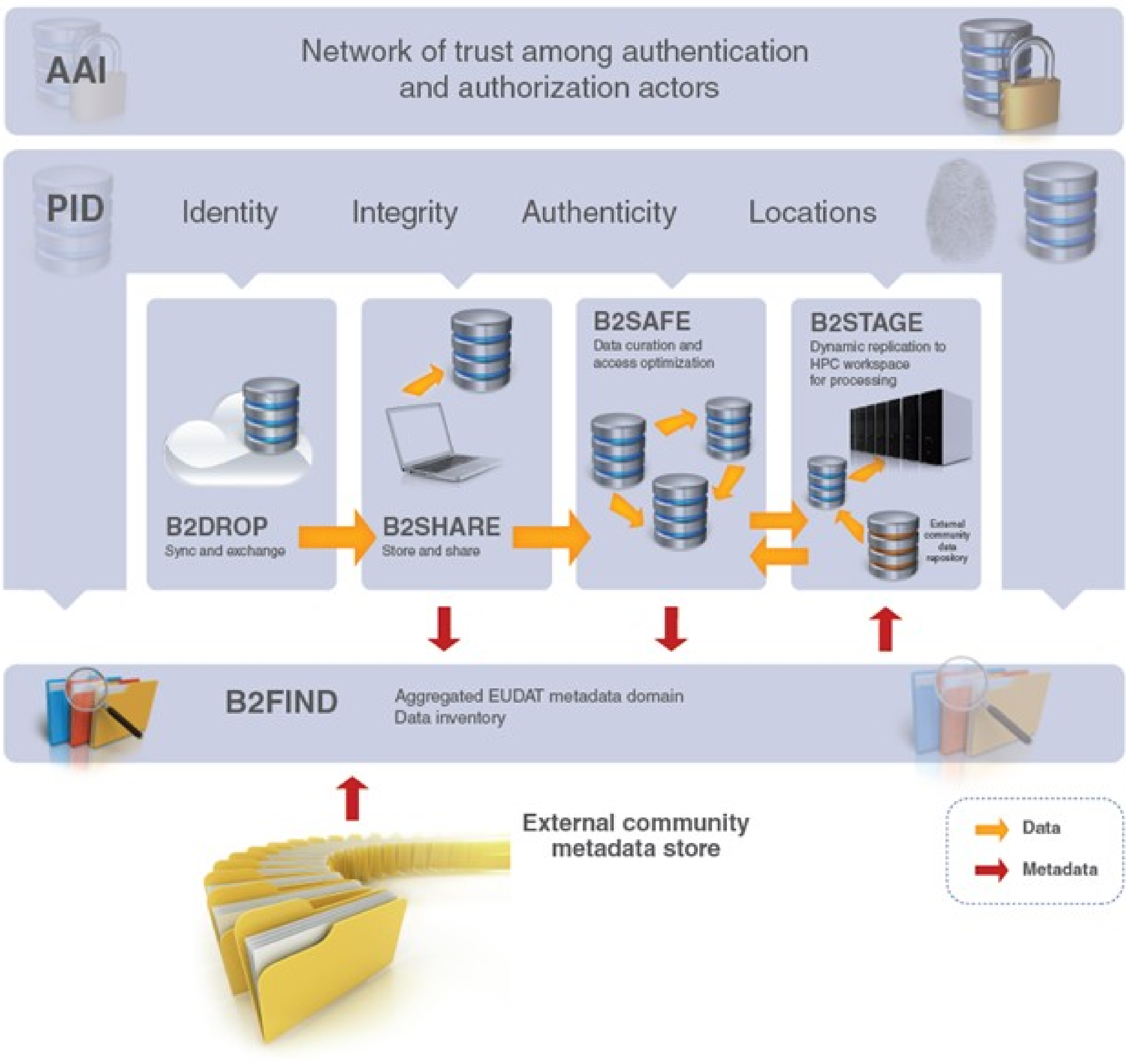
12 September 2014

OpenAIRE

- Supports EC Open Access policy
  - ***Support for beneficiaries and EC***
- Develops and operates an electronic infrastructure
  - ***Free online access to results of EC funded research (and beyond). 10 million publications.***
  - ***Inference and discovery of bibliographic information.***
  - ***Linking publications and data***
  - ***Community collections, cloud storage, DOIs, licensing (Zenodo)***

EUDAT

Services



AAI

Network of trust among authentication and authorization actors

PID

Identity

Integrity

Authenticity

Locations

B2DROP

Sync and exchange

B2SHARE

Store and share

B2SAFE

Data curation and access optimization

B2STAGE

Dynamic replication to HPC workspace for processing

External community data repository

B2FIND

Aggregated EUDAT metadata domain
Data inventory

External community metadata store

Data

Metadata

EUDAT

B2SHARE

Store and Share Research Data

- User-friendly, reliable and trustworthy way for researchers, and citizen scientists to store and share small-scale research data
- Free upload and registration of stable research data
- Data assigned a permanent identifier, which can be retraced to the data owner
- B2SHARE is optimized for researchers who:
  - *do not have adequate facilities for storing research data with metadata,*
  - *cannot guarantee long-term persistence of their locally-stored data, and*
  - *do not have adequate facilities to easily share data, results or ideas with colleagues worldwide*

Trust and repositories

  - European Framework for Audit and Certification of Digital Repositories

- Research Data Alliance Interest Group on Certification
- re3data.org (Registry of Research Data Repositories) is mapping and describing the emerging data repository landscape

Persistent identifiers

- Persistent, trustworthy link
  - **Who wrote/shared this article/dataset?**
  - **What is the dataset supplementary to this article?**
  - **Who cited this article, dataset...?**
- Interoperability between existing resources, linking identifiers across platforms and propagating attribution information
- Robust, persistent mechanism for linking literature and data
- Persistent identifier services
  - **enable search services**
  - **deliver provenance and attribution mechanisms to underpin data exchange**
  - **minting and resolving services for data citation workflows**

Ecoli outbreak 2012

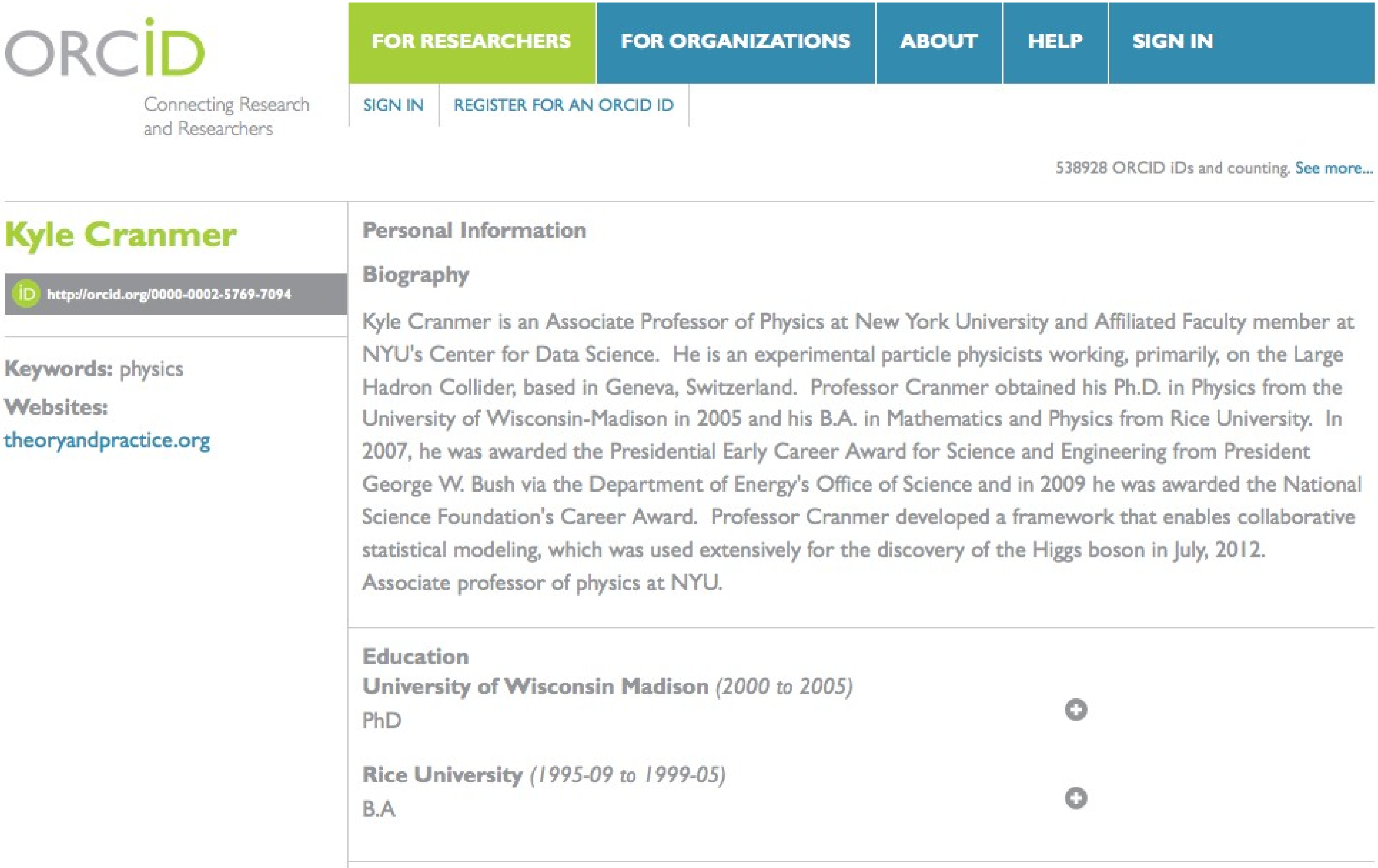
- In the German Ecoli outbreak in 2012 it was the Beijing Genomic institute that made the first genome sequencing of the Ecoli
- Instead of publishing the sequencing as supplement to an article 6 month later, they immediately published the sequence data with a DOI: http://dx.doi.org/10.5524/100001

- Consequently German institutions could identify the source of the outbreak faster thus ending the outbreak

Identifiers for authors and contributors

- Help distinguishing the individual scientist
  - Who wrote, shared, published data, software, papers…
- Similar names, changing names, name variations
- High mobility in a global research world – cannot rely on affiliation, email address etc.
- Trustworthy author identifiers:
  - Unique on a global scale
  - Interoperable, open
  - Across disciplinary boundaries

EC supports ORCID through THOR project



538928 ORCID iDs and counting. See more...

Kyle Cranmer

http://orcid.org/0000-0002-5769-7094

Keywords: physics

Websites:
theoryandpractice.org

Personal Information

Biography

Kyle Cranmer is an Associate Professor of Physics at New York University and Affiliated Faculty member at NYU's Center for Data Science. He is an experimental particle physicists working, primarily, on the Large Hadron Collider, based in Geneva, Switzerland. Professor Cranmer obtained his Ph.D. in Physics from the University of Wisconsin-Madison in 2005 and his B.A. in Mathematics and Physics from Rice University. In 2007, he was awarded the Presidential Early Career Award for Science and Engineering from President George W. Bush via the Department of Energy's Office of Science and in 2009 he was awarded the National Science Foundation's Career Award. Professor Cranmer developed a framework that enables collaborative statistical modeling, which was used extensively for the discovery of the Higgs boson in July, 2012.
Associate professor of physics at NYU.

Education

University of Wisconsin Madison *(2000 to 2005)*
PhD

Rice University *(1995-09 to 1999-05)*
B.A

Looking beyond 2015: open questions

- How to support Horizon 2020 objectives through infrastructures
  - Societal Challenges
  - Innovation, jobs and growth
  - Open Research Data Pilot
- Sustainability of (global) Scientific Data Infrastructure (funders collaboration?)
- Infrastructures that support wide reuse (e.g. TDM), long term preservation
- Etc.

What role for DMPs?

- Mandatory for all projects in Horizon 2020 Open Research Data Pilot
  - Support from OpenAIRE2020 and EUDAT2020 project
- Combination of requirements: EC, other funders, institutions, national, field, …
  - Tool for data sharing
  - Common standard, open source code, machine readable and machine actionable?

Research logic machines

Now **research data** is stored in digital form. Can be processed by "**logic machines**" programmed with complex models able to dig into the data .

Scientist notebooks can now be **linked** to a huge amount of other **data resources** (including scientific papers), **computers** with unprecedented capacity, eventually connected to **global networks.**

New human-machine **interfaces and collaboration software** can be developed to increase the power of these "logic machines".

Digital scholarly record

Publication, data, software, etc. repositories have the potential to become the foundational element of the scholarly record.

+ identifier infrastructures + registries

Registration, linking and validation of **research claims** (using Research Objects?). All higher level **services** and **products** can be added on top.

Possible conditions:

- Authentication and authorisation infrastructure
- Open Access to publications
- Protection of author's rights

take five

5 principles describing the benefits of a global research data infrastructure (G8+O6)

Publicly funded research data is:

Discoverable – IDs, Descriptive Metadata, …

Accessible – Acknowledgment, License, Terms of Use, Intellectual Property, Legal …

Understandable – Semantics, Analysis, Quality, Language translation ….

Manageable – Responsibility, Costs, Preservation …

People (Usable) - Workforce, Cultural, Training, …

The Data Harvest Report

How sharing research data can yield knowledge, jobs and grow

A RDA Europe Report

Useful definitions

Data: digital recorded factual material commonly accepted in the scientific community as necessary to validate research findings

(not include lab notebooks, preliminary analysis, drafts of scientific papers, plans for future research, peer review reports, communication with peers, physical objects, lab specimens)

[c.f. White House Memo on "Increasing Access to the Results of Federally Funded Scientific Research" following OECD definition]

Data infrastructures: services, applications, tools, knowledge and policies for research data to be discoverable, understandable, accessible, preserved and curated… and available 24/7

Thank you!