Research Data Management: A Satellite Workshop to the 67th Annual
Meeting of the American Crystallographic Association
Nouvelle Orléans, United States of America
26 May 2017

Open Science, FAIR Data and Data Standards

Marshall Ma, University of Idaho; Co-Chair, CODATA TG on Coordinating Data Standards amongst Scientific Unions, *and*

Simon Hodson, Executive Director, CODATA

www.codata.org

CODATA Prospectus:

https://doi.org/10.5281/zenodo.165830

Principles, Policies and Practice

Current Best Practice for Research Data Management Policies

A Memo for the Danish e-Infrastructure Cooperation and the Danish Digital Library

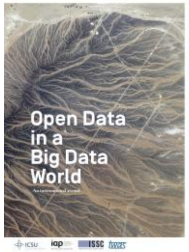
Simon Hodson and Laura Molloy

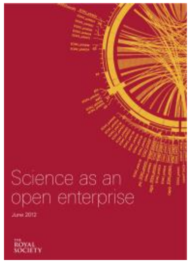
May 2014

DC1 Data Citation Principles

INTERNATIONAL COUNCIL FOR SCIENCE

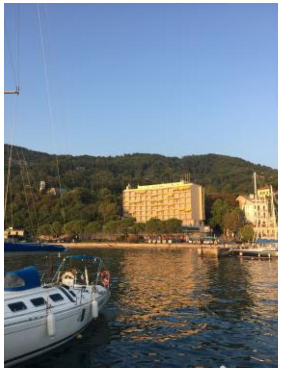
The Value of Open Data Sharing


Open Data in a Big Data World


Science as an open enterprise

Capacity Building

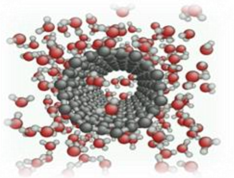
中國科學院
CHINESE ACADEMY OF SCIENCES

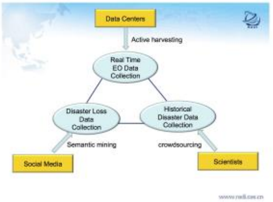


UNESCO

The Abdus Salam International Centre for Theoretical Physics

Frontiers of Data Science





Workshop on Big Data for International Scientific Programmes: Challenges and Opportunities
8-9 June 2014 Beijing · China
CODATA
WORLD DATA SYSTEM
futurearth
IRDR
RDA
GEO
北京国际会议中心
Beijing International Convention Centre

]u[ubiquity press open scholarship

ISSN 1234-1234 (online)
Volume 1 (2015)
DATA SCIENCE JOURNAL
Editor in Chief: Sarah Callaghan

Data Science Journal

INTERNATIONAL DATA WEEK 2016
WWW.INTERNATIONALDATAWEEK.ORG
Organized by:

SciDataCon 2016
Advancing the Frontiers of Data in Research
11–13 September, Denver, Colorado, USA

International Conference
CODATA 2017

"GLOBAL CHALLENGES and DATA-DRIVEN SCIENCES"

CODATA 2017, Saint Petersburg 8-13 Oct 2017

http://codata2017.gcras.ru/

Why Open Science / FAIR Data?

- **Good scientific practice depends on communicating the evidence.**
  - Open research data are essential for reproducibility, self-correction.
  - Academic publishing has not kept up with age of digital data.
  - Danger of an replication / evidence / credibility gap.
  - Boulton: to fail to communicate the data that supports scientific assertions is malpractice
- **Open data practices have transformed certain areas of research.**
  - Genomics and related biomedical sciences; crystallography; astronomy; areas of earth systems science; various disciplines using remote sensing data…
  - **FAIR data helps use of data at scale, by machines, harnessing technological potential.**
  - Research data often have considerable potential for reuse, reinterpretation, use in different studies.
- **Open data foster innovation and accelerate scientific discovery through reuse of data within and outside the academic system.**
  - Research data produced by publicly funded research are a public asset.

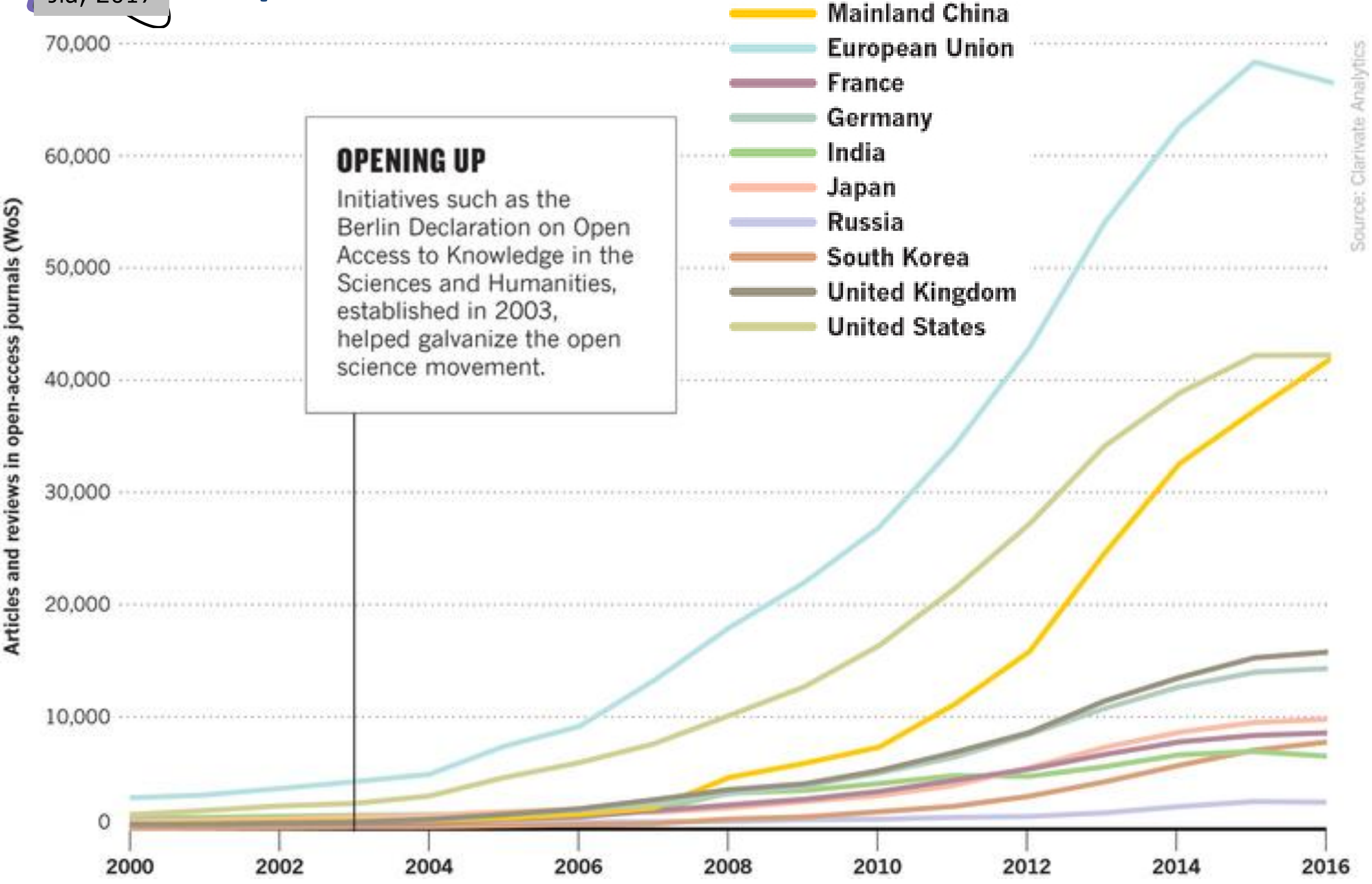
Riding the wave
How Europe can gain from the rising tide of s
Final report of the High Level Expert Group on Scientific D
A submission to the European Commission
October 2010
Science as an open enterprise
June 2012
THE ROYAL SOCIETY
A WORLD THAT COUNTS
MOBILISING THE DATA REVOLUTION FOR SUSTAINABLE DEVELOPMENT
GODAN
ODI
How can we improve agriculture, food and nutrition with open data?
UK20150528
Open Data in a Big Data World
An international accord
ICSU
iap
ISSC
twas

Policy Push for Open Research Data

- The three Bs (Budapest, Berlin and Bethesda) and Open Access, 2002-3
- OECD Principles and Guidelines on Access to Research Data, 2004, 2007
- UK Funder Data Policies, from 2001, but accelerates from 2009
- NSF Data Management Plan Requirements, 2010
- Royal Society Report ‘Science as an Open Enterprise’, 2012
- OSTP Memo ‘Increasing Access to the Results of Federally Funded Scientific Research’, Feb 2013
- G8 Science Ministers Statement, June 2013
- G8 Open Data Charter and Technical Appendix, June 2013
- EC H2020 Open Data Policy Pilot, 2014; Adoption of FAIR Data Principles, 2017.
- Science International Accord on Open Data in a Big Data World, Dec 2015: http://bit.ly/opendata-bigdata

The three Bs (Budapest, Berlin and Bethesda) and Open Access, 2002-3

Jia, 2017



OPENING UP

Initiatives such as the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, established in 2003, helped galvanize the open science movement.

Source: Clarivate Analytics

Data policies of major research funders in USA

Dietrich et al. 2012

Research Funders	National Science Foundation (NSF)	NSF Basic Research to Enable Agricultural Development (BREAD)	NSF Division of Earth Sciences (EAR)	NSF Division of Ocean Sciences	NSF Integrated Ocean Drilling Program	NSF Ocean Acidification Reaserach	NSF Office of Polar Programs	NSF Engineering Directorate	NSF Social Behavioral and Economic Sciences	National Institutes of Health (NIH)	NIH - Genome-Wide Association Studies (GWAS)	NIH - National Human Genome Research Institute	United States Department of Agriculture (USDA)	United States Department of Energy (DOE)	DOE Atmospheric Radiation Measurement Program (ARM)	United States Department of Education (DoEd)	United States Environmental Protection Agency (EPA)	United States Agency for International Development (USAID)	National Aeronautics and Space Administration (NASA) - Heliophysics	National Aeronautics and Space Administration (NASA) – Earth Sciences	Office of Naval Research (ONR)	Office of Naval Research Policy for In Situ Ocean Data (ONR)	National Oceanographic and Atmospheric Administration (NOAA) Climate Observations and Monitoring (COM)	National Oceanographic and Atmospheric Administration (NOAA) Coastal Ocean Program (COP)	American Heart Association	Sloan Foundation
Open Access to Publications	S	S	R	O	O	O	O	O	O	R	R	R	S	O	O	O	O	O	O	R	O	O	O	R	O	O
Publication Repository Specified	O	O	R	O	R	O	O	O	O	R	R	R	O	O	O	O	O	O	O	O	O	O	O	R	O	O
Publication Repository Supported	O	O	O	O	O	O	O	O	O	R	R	R	O	O	S	O	O	O	O	O	O	O	O	O	O	O
Organizational Data Policy	R	R	R	R	R	R	R	R	R	R	R	R	R	S	R	R	O	R	R	R	O	R	O	R	O	O
Data Plan in Proposals	R	R	R	R	R	R	R	R	R	R	R	R	R	O	R	O	O	O	R	R	O	O	R	O	O	O
Data Timeframe	S	R	R	R	R	R	R	R	R	R	R	R	S	O	R	O	O	O	R	R	O	R	O	R	O	O
Data Access	R	R	R	R	R	R	R	R	R	R	R	R	R	O	R	O	O	R	R	R	O	O	R	R	O	O
Data Embargo	S	R	S	O	S	O	O	S	O	S	S	S	S	S	S	O	O	O	O	O	O	O	O	O	O	O
Data Preservation	S	O	R	R	O	R	O	R	O	R	R	R	S	O	R	S	O	R	R	R	O	R	R	R	O	O
Data Standards	S	R	O	O	O	R	R	R	R	O	O	O	S	O	R	O	O	O	R	O	O	R	O	O	O	O
Metadata Standards	S	O	O	S	O	R	R	R	O	R	O	R	S	O	R	O	O	O	R	O	O	O	O	O	O	O
Compliance	O	O	R	R	R	O	R	R	O	R	O	R	O	O	O	O	O	O	O	R	O	O	O	O	O	O
Data Center Specified	O	R	R	R	O	O	R	O	R	R	R	R	O	O	R	O	O	R	R	R	O	R	R	R	O	O
Data Center Supported	O	O	O	S	O	O	O	O	O	O	R	O	O	O	R	O	O	O	R	R	O	O	R	R	O	O
Funding	O	R	O	O	O	R	S	O	O	R	R	R	O	O	R	O	O	O	O	R	O	O	R	O	O	O
Scope	O	R	R	R	R	R	R	O	O	R	O	R	O	O	R	O	O	O	R	R	O	R	R	R	O	O
Guidance	O	O	O	O	O	S	O	S	S	O	R	R	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Policy Date	2010		2002	1988, 1994, 2003	2009		1998			2003	2007, 2008	2008	2008			2008		2010	2009	1990s, 2000s	2010	1999	2011	2002		

The Case for Open Data in a Big Data World

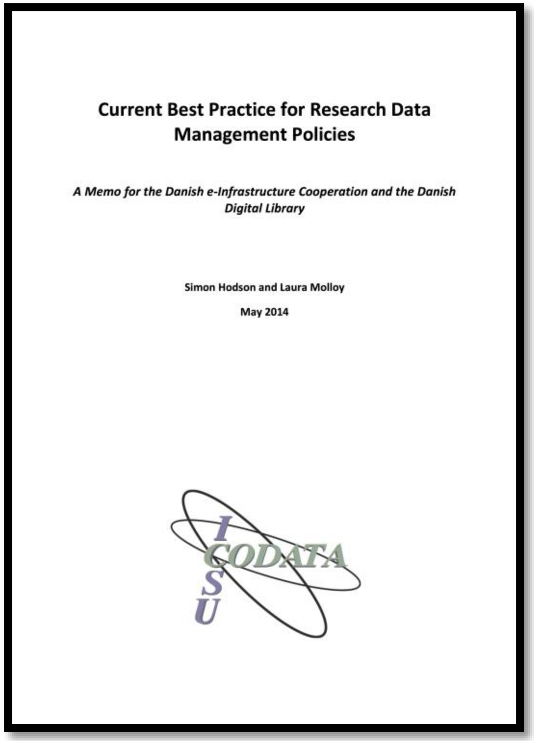
- **Science International Accord on Open Data in a Big Data World: http://www.science-international.org/**
- Presents a powerful case that the profound transformations mean that data should be:
    - Open by default
    - Intelligently open
- Supported by four major international science organisations.
- **Lays out a framework of principles, responsibilities and enabling practices for how the vision of Open Data in a Big Data World can be achieved.**
- Campaign for endorsements: over 100 organisations so far. **Please consider endorsing the Accord.**
- Translations: Chinese, Russian, Polish, Spanish, French.
- **IUCr Position Paper in response: http://www.iucr.org/iucr/open-data**

Open Data in a Big Data World

An international accord

Resources: Current Best Practice for Research Data Management Policies

- **Expert report commissioned by CODATA member.**
- Provides comprehensive summary of best practice in funder data policies.
- Identifies key elements to be addressed:
    1. Summary of policy drivers
    2. Intelligent openness
    3. **Limits of openness**
    4. **Definition of research data**
    5. **Define data in scope**
    6. **Criteria for selection**
    7. Summary of responsibilities
    8. Infrastructure and costs
    9. DMP requirements
    10. Enabling discovery and reuse
    11. Recognition and reward
    12. Reporting requirements, compliance monitoring
- Zenodo: http://dx.doi.org/10.5281/zenodo.27872
- See also **RECODE Report**, Annex on Policy Development: http://recodeproject.eu/
- LEARN Project Toolkit: http://learn-rdm.eu/en/about/
- FOSTER Knowledge Base on Open Science:



Current Best Practice for Research Data Management Policies

A Memo for the Danish e-Infrastructure Cooperation and the Danish Digital Library

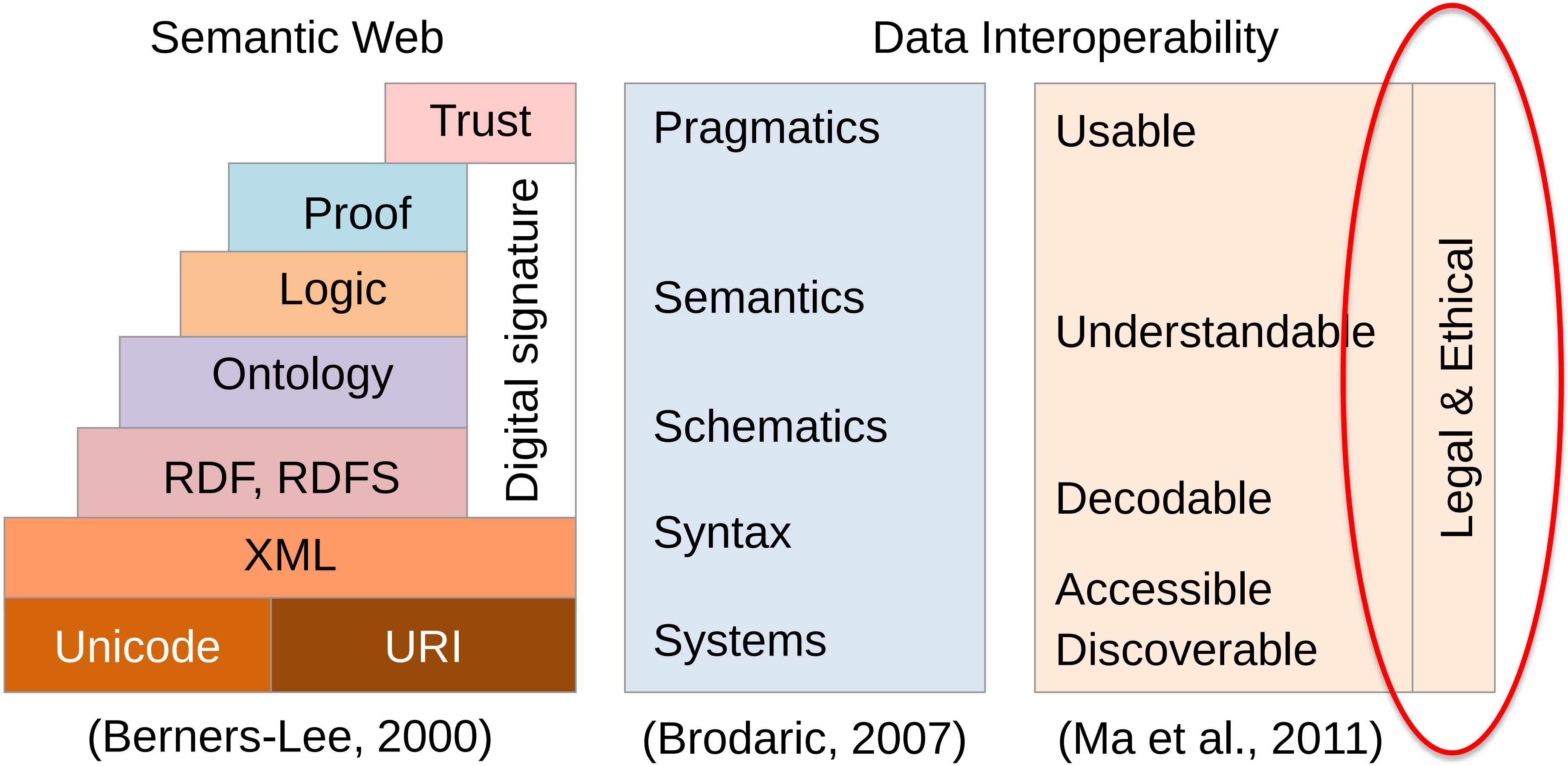
Simon Hodson and Laura Molloy

May 2014

Socio-technological system of open data and data interoperability

Semantic Web

Data Interoperability



Trust

Proof

Logic

Ontology

RDF, RDFS

XML

Unicode

URI

Digital signature

Pragmatics

Semantics

Schematics

Syntax

Systems

Usable

Understandable

Decodable

Accessible

Discoverable

Legal & Ethical

(Berners-Lee, 2000)

(Brodaric, 2007)

(Ma et al., 2011)

Open data vs. Closed data

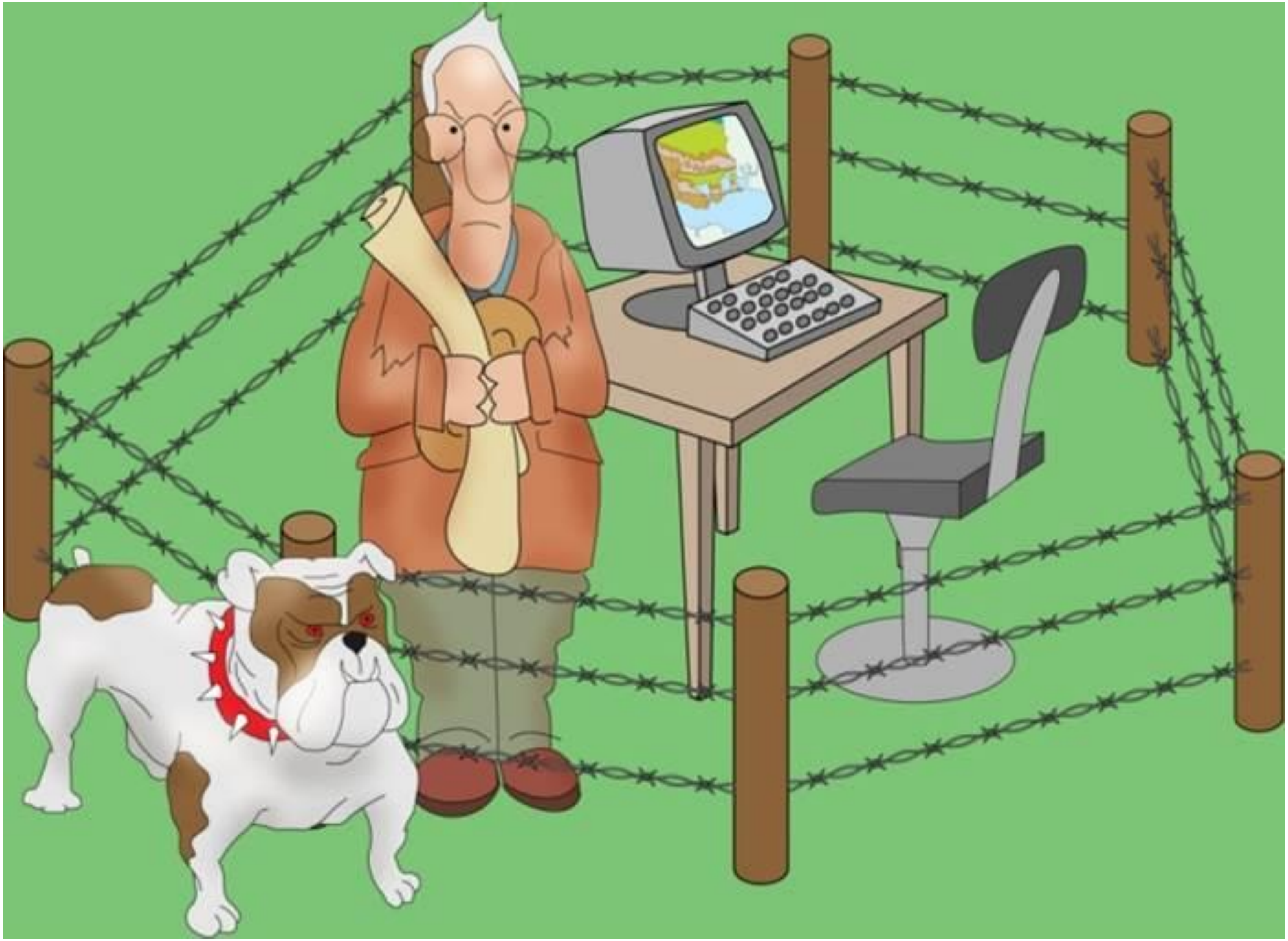
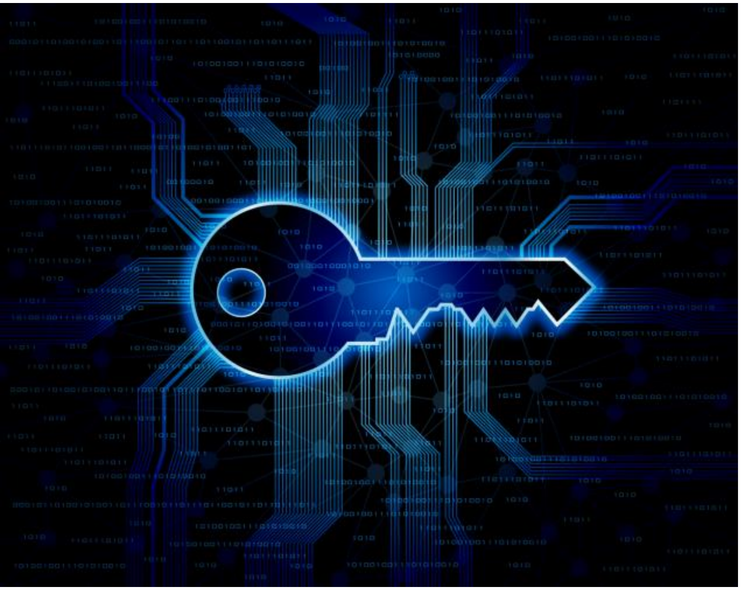


Image courtesy of British Geological Survey

Boundaries of Open



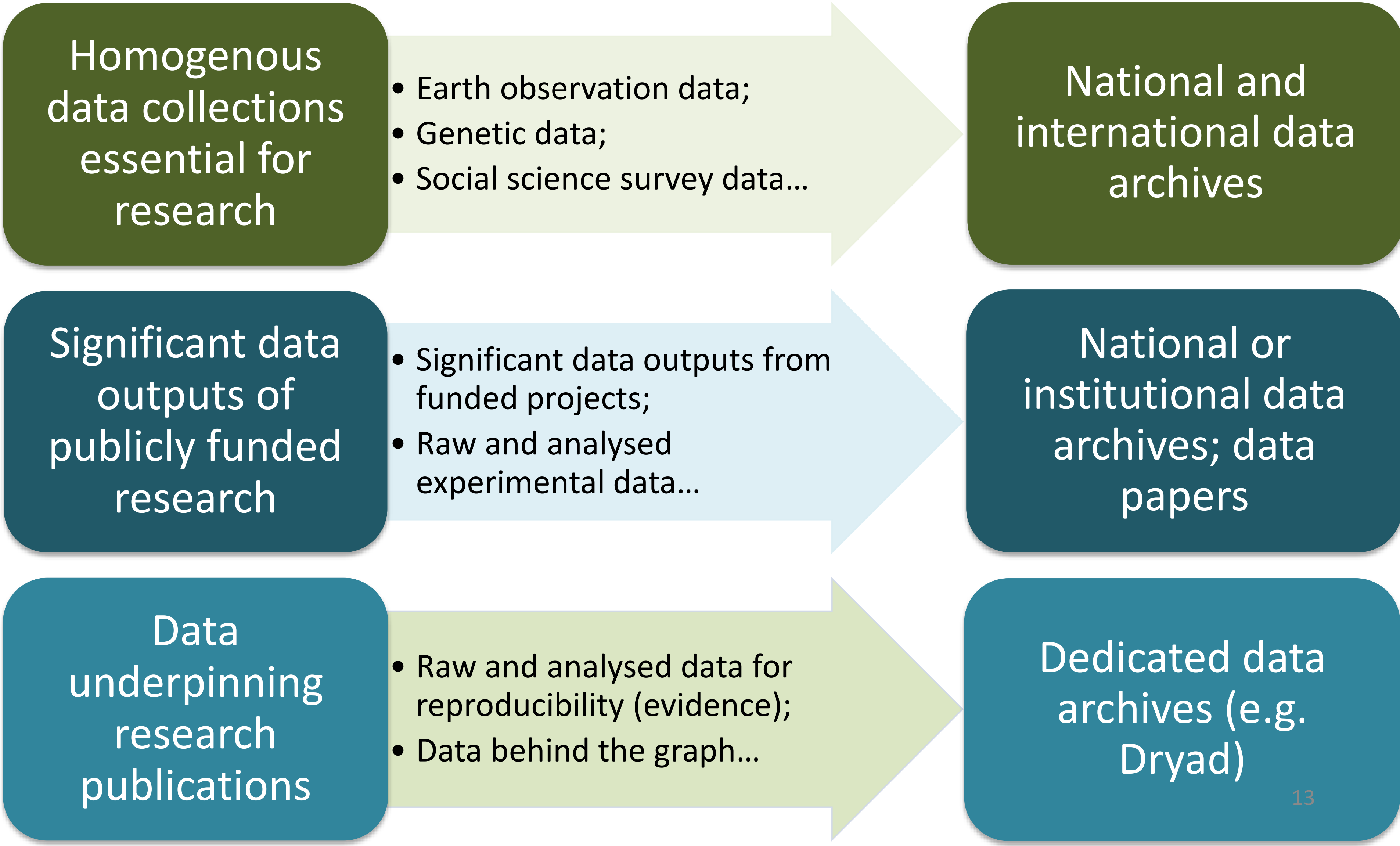
- For data created with public funds or where there is a strong demonstrable public interest, **Open should be the default.**
- **As Open as Possible as Closed as Necessary.**
- **Proportionate** exceptions for:
  - Legitimate **commercial** interests (sectoral variation)
  - **Privacy** (‘safe data’ vs Open data – the anonymisation problem)
  - **Public interest** (e.g. endangered species, archaeological sites)
  - **Safety, security** and dual use (impacts contentious)
- All these boundaries are fuzzy and need to be understood better!
- **There is a need to evolve policies, practices and ethics around closed, shared, and open data.**

Implementation Guidelines for the Legal Interoperability of Research Data

1. **Facilitate the lawful access to and reuse of research data.**
2. **Determine the rights to and responsibilities for the data.**
3. **Balance the legal interests.**
4. **State the rights transparently and clearly.**
5. **Promote the harmonization of rights in research data.**
6. **Provide proper attribution and credit for research data.**

- **Joint CODATA-RDA Interest Group on Legal Interoperability.**
- Builds on work done in the context of the GEO Data Sharing Working Group.
- Set of principles to help ensure the fewest possible legal barriers relating to IP to sharing research data.
- Implementation guidelines offer high level guidance on steps to take to reduce legal barriers to data reuse.
- Result of lengthy consideration by the IG and two strenuous rounds of peer review.
- **Final version of the guidelines:** https://doi.org/10.5281/zenodo.162241

Where should research data go?



Homogenous data collections essential for research

- Earth observation data;
- Genetic data;
- Social science survey data…

National and international data archives

Significant data outputs of publicly funded research

- Significant data outputs from funded projects;
- Raw and analysed experimental data…

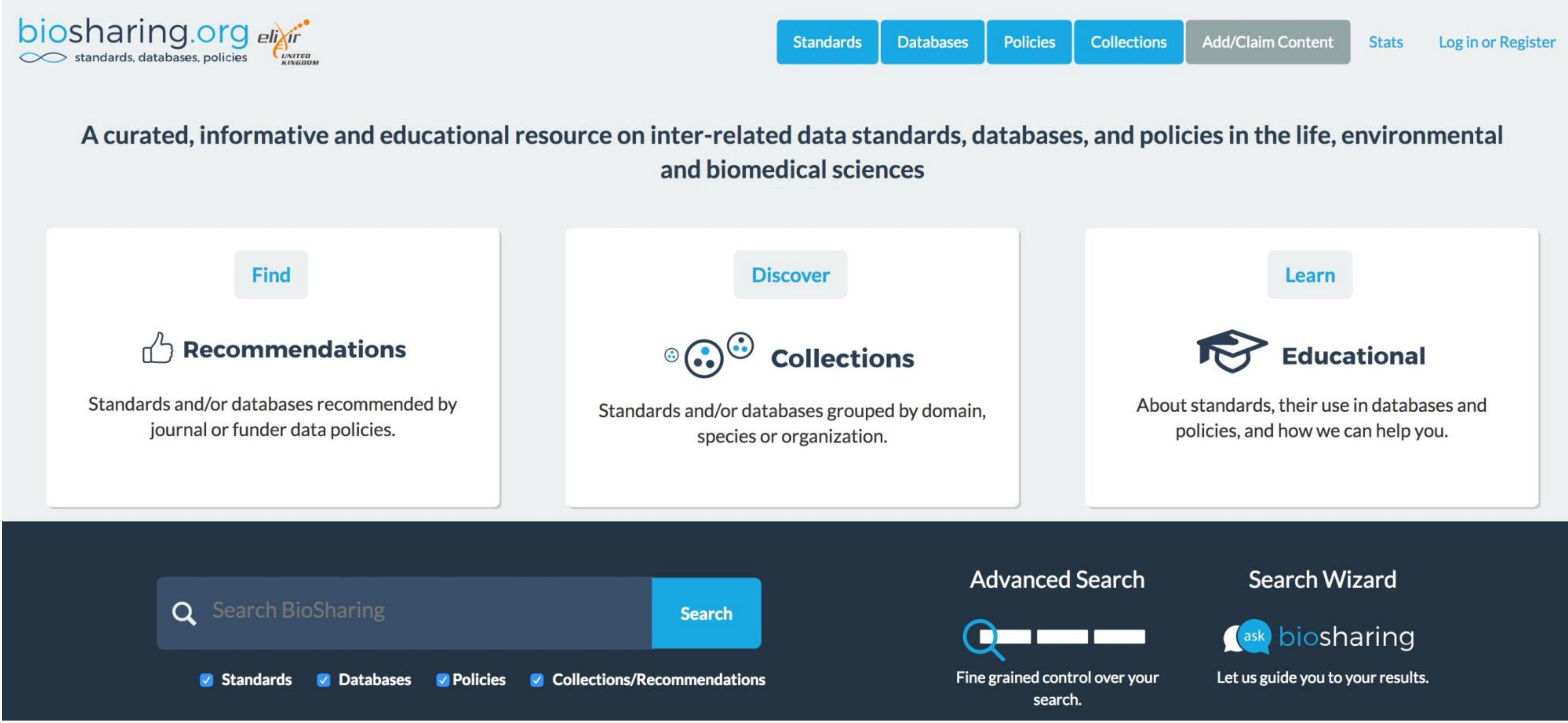
National or institutional data archives; data papers

Data underpinning research publications

- Raw and analysed data for reproducibility (evidence);
- Data behind the graph…

Dedicated data archives (e.g. Dryad)

BioSharing to FAIRsharing



A curated, informative and educational resource on inter-related data standards, databases, and policies in the life, environmental and biomedical sciences

Recommendations

Standards and/or databases recommended by journal or funder data policies.

Collections

Standards and/or databases grouped by domain, species or organization.

Educational

About standards, their use in databases and policies, and how we can help you.

Fine grained control over your search.

Let us guide you to your results.

Emerging Policy Consensus? FAIR Data

- **FAIR Data** (see original guiding principles at https://www.force11.org/node/6062
  - **Findable:** have sufficiently rich metadata and a unique and persistent identifier.
  - **Accessible:** retrievable by humans and machines through a standard protocol; open and free; authentication and authorization where necessary.
  - **Interoperable:** metadata use a ‘formal, accessible, shared, and broadly applicable language for knowledge representation’.
  - **Reusable:** metadata provide rich and accurate information; clear usage license; detailed provenance.
- FAIR Data now at the heart of H2020 policy, European Open Science Cloud etc.
  - **Under the revised version of the 2017 work programme, the Open Research Data pilot has been extended to cover all the thematic areas of Horizon 2020.**
- Current EC Guidance at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf and http://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf
- **European Commission Expert Group (chaired by Simon Hodson, CODATA; Sarah Jones, DCC, Rapporteur) producing implementation guidelines for FAIR Data for EC Funded Programmes: draft report end 2017, final report March 2018: http://bit.ly/FAIRdata-EG**

FAIR Guiding Principles (1)

- **To be Findable:**
  - F1. (meta)data are assigned a globally unique and persistent **identifier**
  - F2. data are described with rich metadata (defined by R1 below)
  - F3. metadata clearly and explicitly include the **identifier** of the data it describes
  - F4. (meta)data are registered or indexed in a searchable resource
- **To be Accessible:**
  - A1. (meta)data are retrievable by their **identifier** using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  - A2. metadata are accessible, even when the data are no longer available

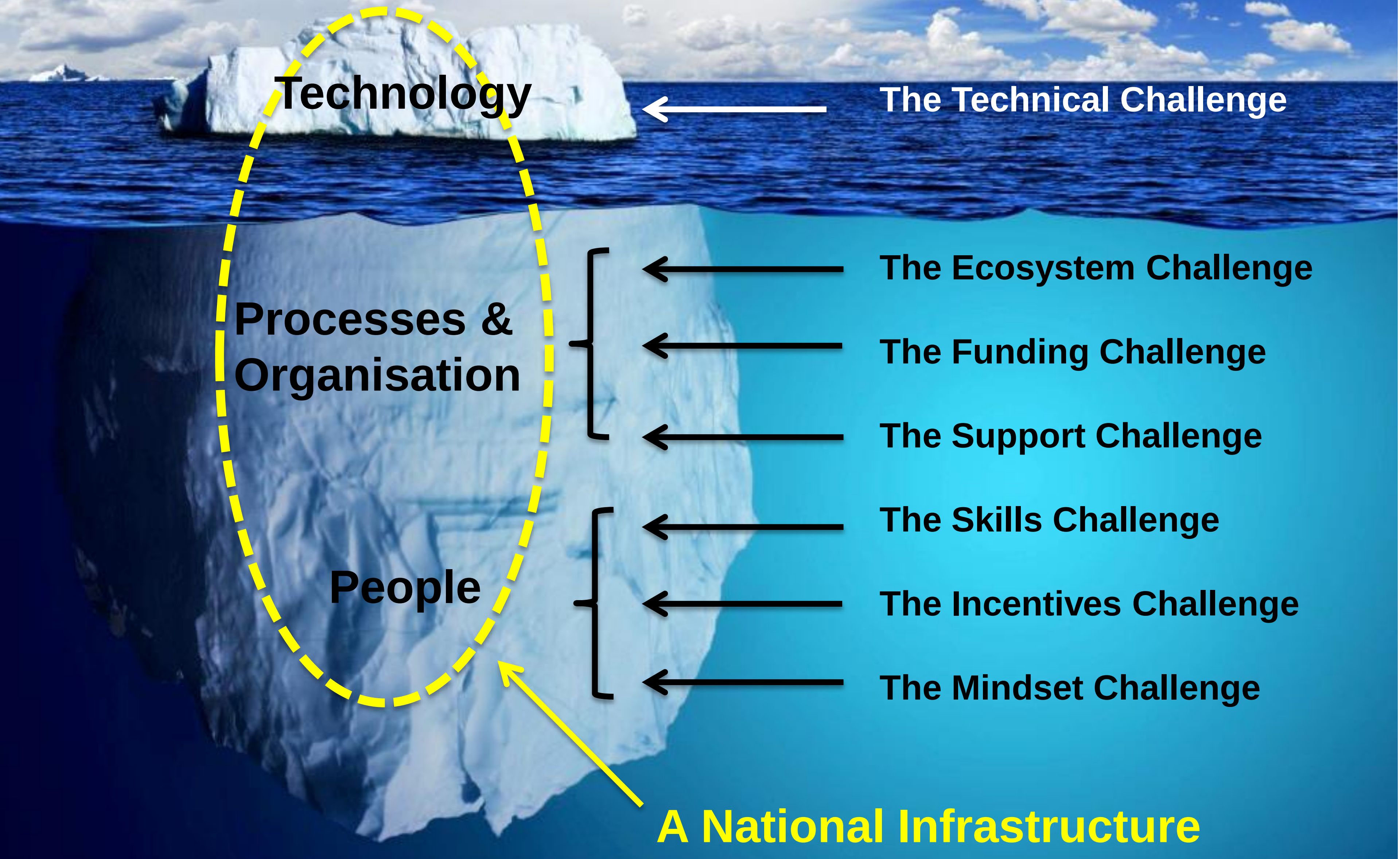
(Mons, B., et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, http://dx.doi.org/10.1038/sdata.2016.18)

FAIR Guiding Principles (2)

- **To be Interoperable:**
  - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (meta)data use vocabularies that follow FAIR principles
  - I3. (meta)data include qualified references to other (meta)data
- **To be Reusable:**
  - R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

(Mons, B., et al., The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, http://dx.doi.org/10.1038/sdata.2016.18)

The Open Data Iceberg



Technology

Processes & Organisation

People

The Technical Challenge

The Ecosystem Challenge

The Funding Challenge

The Support Challenge

The Skills Challenge

The Incentives Challenge

The Mindset Challenge

A National Infrastructure

Geoffrey Boulton (CODATA) - developed from an idea by Deetjen, U., E. T. Meyer and R. Schroeder (2015). *OECD Digital Economy Papers*, No. 246, OECD Publishing.

Data is difficult: motivations and reward

- FAIR data is an attempt to communicate principles which make data more reusable by both humans and machines.
- **Opens the prospect of analysing data at scale, across the web.** For this opportunity to be seized data needs to be well-described, unambiguously identified, etc.
- Data analysis at greater scale; integration of data from diverse sources; more effective data products and reporting (e.g. SDGs, Sendai etc); use of data by practitioners (farmers, hydrologists, economists, etc) and entrepreneurs.
- Data description, definitions and ontologies, data management require significant effort…
- Increasing calls from research funders and research communities for that effort to be made and facilitated.
- Researchers' and institutions' reputation will increasingly rely on data as well as other contributions – data is one of the assets which build reputation and enable collaborations.
- When we communicate our knowledge and research outputs open for discussion, increasingly that means data too.
- **Data is becoming increasingly integrated with the process of scholarly communications.**

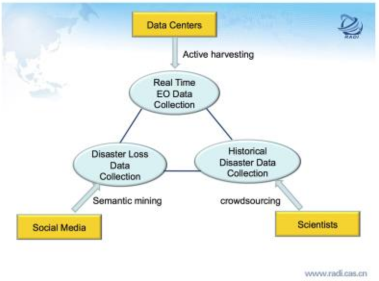
CODATA TG
Coordinating Data Standards amongst Scientific Unions
- established in 2016 -

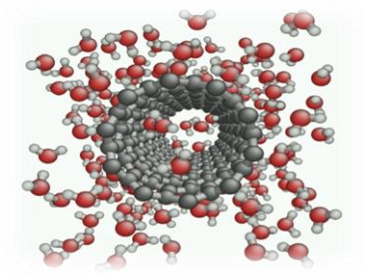
- Identify data and information commission or WG in Science Unions
- Use the CODATA web site and social media to raise awareness of standards endorsed by the individual unions
- Provide a web-accessible page that provides links to repositories for data models, information standards, vocabularies, ontologies, etc., for each of the unions
- Determine a broad ‘maturity model’ for scientific standards adapted from the 5 star Open Data model and the AGU Data Maturity Framework
- Provide best practice examples for the development, governance and application of the required standards
- Provide guidelines to the scientific community for the need to adhere to these standards and promote the benefits of adherence to standards

http://www.codata.org/task-groups/coordinating-data-standards

Commission on Data Standards for Science

- Major transdisciplinary research issues depend on the integration of data and information from different sources.
- Fundamental importance of agreed vocabularies and standards.
  - Importance of integration of social science, geospatial and other data
  - Essential to effective interface of science and monitoring (e.g. Sendai, SDGs)
  - LOD for Disaster Research, Nanomaterials Uniform Description System
- Huge opportunities but significant challenges.
- The ICSU and ISSC, any merged Council, and international scientific unions could have a major role to play to encourage and accelerate these developments.
- **‘Inter-Union Workshop on 21st Century Scientific and Technical Data Developing a roadmap for data integration’, Paris, 19-20 June**.
- Larger follow-up workshop later in the year.
- Vision of a decadal initiative to advance science through integration of data and information.





INTERNATIONAL COUNCIL FOR SCIENCE

CODATA 2017

http://codata2017.gcras.ru/

http://conference.codata.org/2017/

RSF Russian Science Foundation

INTERNATIONAL COUNCIL FOR SCIENCE

RESEARCH & ENGINEERING CORPORATION MEKHANOBR TEKHNIKA

MISIS

IIASA

ISSC

INTERNATIONAL CONFERENCE

GLOBAL CHALLENGES AND DATA-DRIVEN SCIENCE

Saint Petersburg

08 October 2017 – 13 October 2017

CODATA 2017: Global Challenges and Data Driven Science

Major conference themes:

1. Achievements in Data Driven Science, in all research disciplines
2. Earth Observations Data and the Earth's System
3. Data and Disaster Risk Research
4. Data Driven and Sustainable Cities
5. Big Data in Scientific and Commercial Sectors
6. Data Analysis, Event Recognition and Applications
7. National and International Data Services
8. Research Data Services in Universities
9. Coordination of Data Standards and Interoperability
10. FAIR Data and the Limits of Open Data
11. Metrology, Reference Data and Monitoring Data

Call for Sessions and Papers

- **Deadline is 30 June 2017**
- Participants may submit session proposals, with proposed papers, or papers against conference themes.
- Encouraged to submit papers for a special collection in the Data Science Journal

INTERNATIONAL
COUNCIL
FOR SCIENCE

Thank you for your attention!

Marshall Ma
University of Idaho

Simon Hodson
Executive Director CODATA

www.codata.org
http://lists.codata.org/mailman/listinfo/codata-international_lists.codata.org
Email: simon@codata.org
Twitter: @simonhodson99
Tel (Office): +33 1 45 25 04 96 | Tel (Cell): +33 6 86 30 42 59

CODATA (ICSU Committee on Data for Science and Technology), 5 rue Auguste Vacquerie, 75016 Paris, FRANCE