OpenData.CH/2017
Hochschule Luzern
27 June 2017

Open and FAIR Research Data: how do we get there?

Simon Hodson, Executive Director, CODATA
www.codata.org

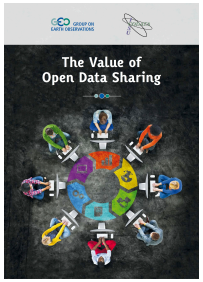
CODATA Prospectus:

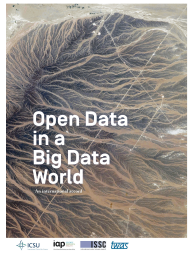
https://doi.org/10.5281/zenodo.165830

Principles, Policies and Practice

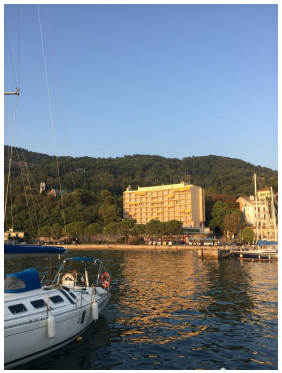
DC1 Data Citation Principles

INTERNATIONAL COUNCIL FOR SCIENCE

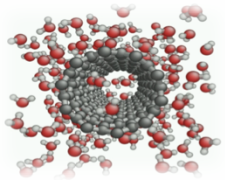


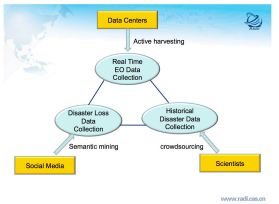


Capacity Building



Frontiers of Data Science





ubiquity press open scholarship

Data Science Journal

INTERNATIONAL DATA WEEK 2016
WWW.INTERNATIONALDATAWEEK.ORG

CODATA 2017, Saint Petersburg 8-13 Oct 2017

http://codata2017.gcras.ru/

Why Open Science / FAIR Data?

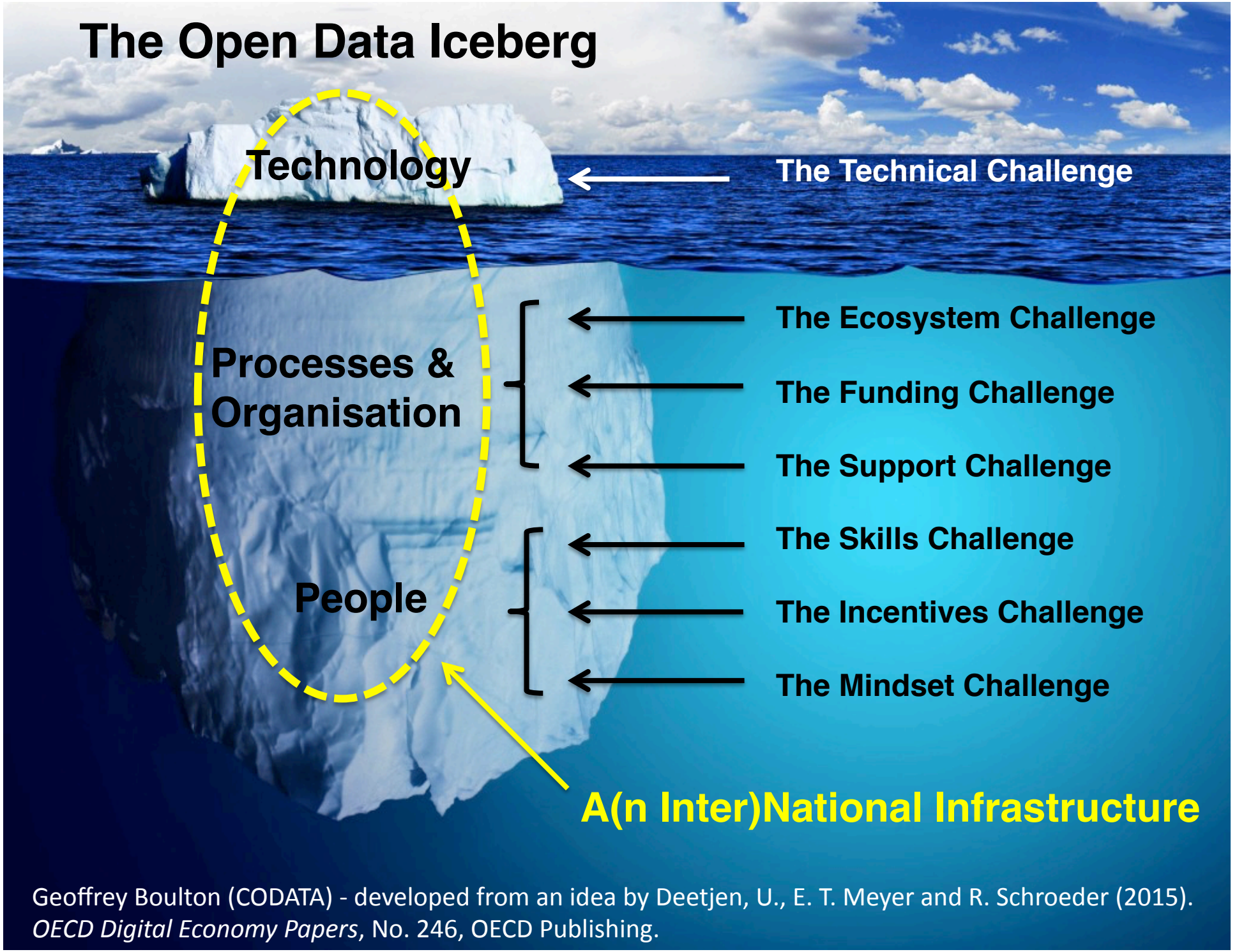
- **Good scientific practice depends on communicating the evidence.**
  - Open research data are essential for reproducibility, self-correction.
  - Academic publishing has not kept up with age of digital data.
  - Danger of an replication / evidence / credibility gap.
  - Boulton: to fail to communicate 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.

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 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/**
- Supported by four major international science organisations.
- Presents a powerful case that the profound transformations mean that data should be:
  - Open by default
  - Intelligently open
- **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 150 organisations so far.
- **Please consider endorsing the Accord: http://www.science-international.org/#endorse**



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.

Open and FAIR Data: how do we get there?

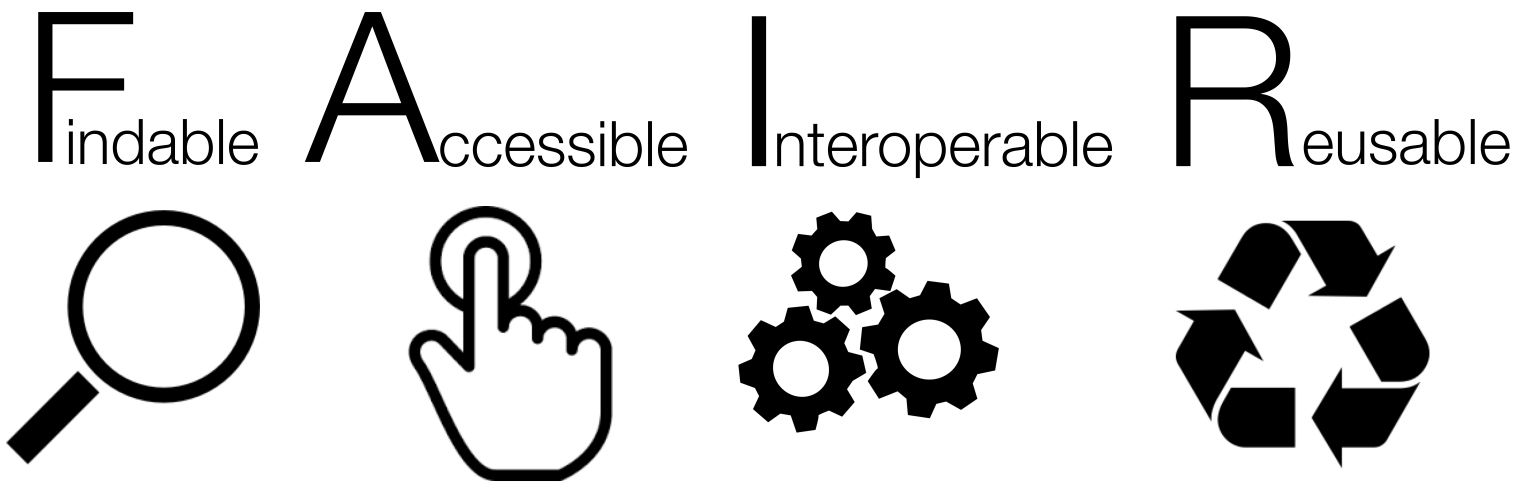
- Clarify the boundaries of Open for research data.
- Refine and improve understanding of FAIR data.
- Work with **and across** disciplines on standards and vocabularies.
- Invest in sustainable data infrastructure (including repositories, stewardship, standards), and develop appropriate business models for sustainability.
- Incorporate research data in the process of scholarly communication and ensure that researchers, research groups and institutions receive adequate reward and recognition for their efforts.

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.**

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 by default; 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.



Emerging Policy Consensus? FAIR Data

- Builds on previous definitions: e.g. 'Intelligent Openness' or G8 Science Ministers' Statement (discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards).
- 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://bit.ly/EC_H2020_RDM_Guidance and http://bit.ly/EC_H2020_OpenData_Infographic
- **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**
- **Call for suggestions and contributions on implementing the FAIR data principles: http://bit.ly/FAIR_Data_Consultation**

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)

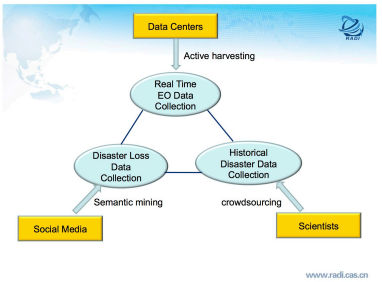
FAIR Guiding Principles (2)

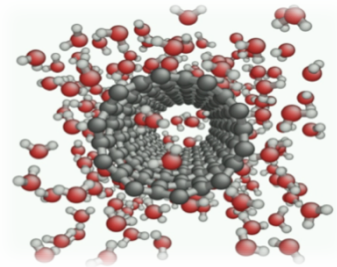
- **To be Interoperable:**
  - I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (meta)data use vocabularies that follow FAIR principles
  - I3. (meta)data include qualified references to other (meta)data
- **To be Reusable:**
  - R1. 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)

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.
  - Fundamental to integration of social science, geospatial and other data
  - Essential to effective interface of science and monitoring (e.g. Sendai, SDGs, sustainable cities)
  - 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: http://bit.ly/codata_standards_workshop**
- Larger follow-up workshop later in the year.
- Vision of a decadal initiative to advance science through integration of data and information.





CODATA WG on Description of Nanomaterials

Convene ISUs, International Stakeholders and data experts → Form Working Group → Draft Framework for Description of Nanomaterials → Refine/validate in FP7 Future Nano Needs Project

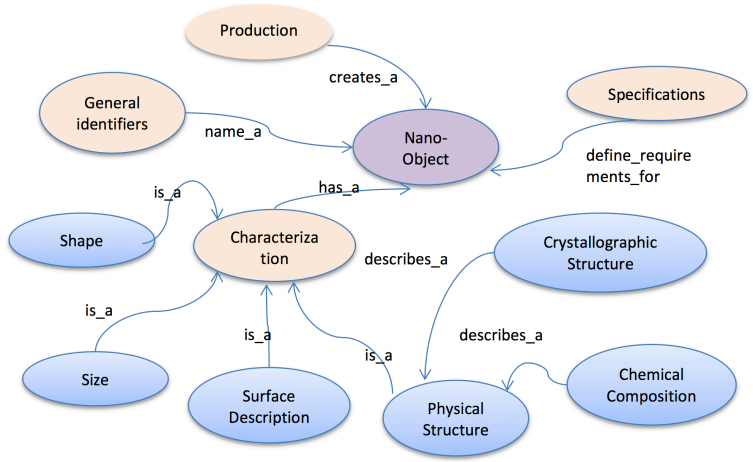


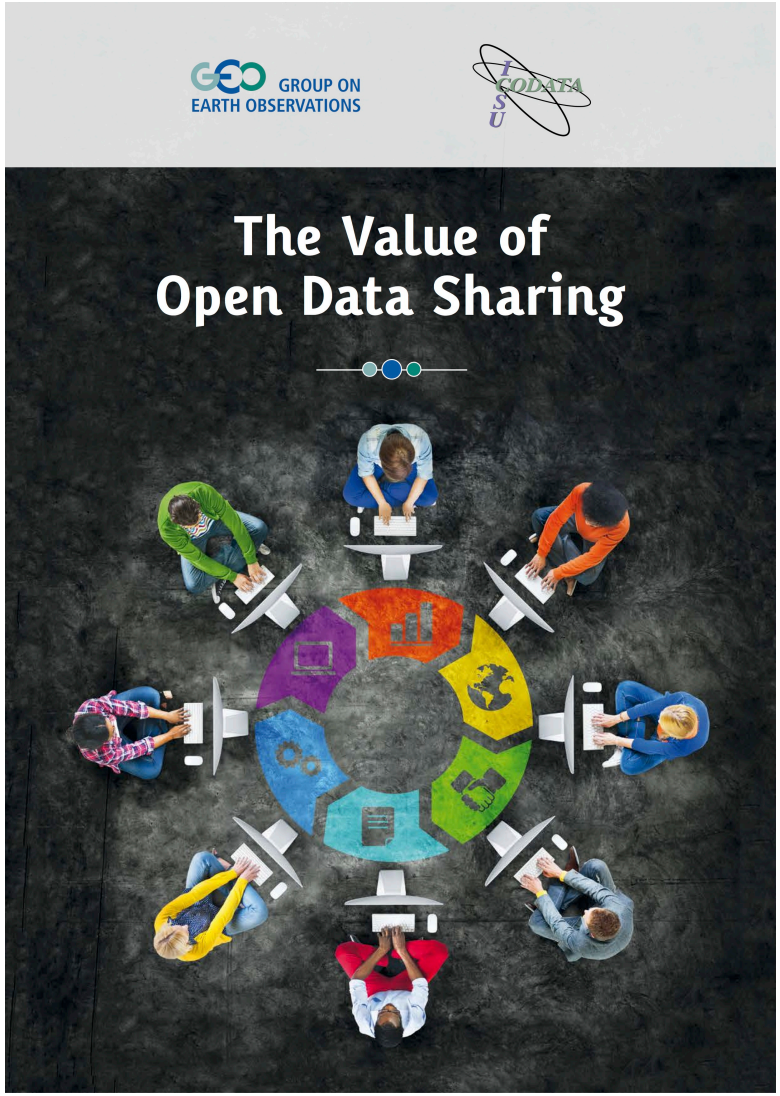
Figure 4. Information categories for describing an individual nano-object

CODATA WG on the Description of Nanomaterials:
http://www.codata.org/nanomaterials

Uniform Description System v.02, May 2016:
http://dx.doi.org/10.5281/zenodo.56720

Future Nano Needs Project:
http://www.futurenanoneeds.eu/

The Value of Open Data Sharing



- Report by CODATA for GEO, the Group on Earth Observation.
- Provides a concise, accessible, high level synthesis of key arguments and evidence of the benefits and value of open data sharing.
- Particular, but not exclusive, reference to Earth Observation data.
- Benefits in the areas of:
  - Economic Benefits
  - Social Welfare Benefits
  - Research and Innovation Opportunities
  - Education
  - Governance
- Available at http://dx.doi.org/10.5281/zenodo.33830
- **GEO DSWG is building on this work with further examples: would be valuable to work with this community.**

The Challenge: Business Models for Sustainable Data Repositories

- Research increasingly relies on digital repositories, curated databases and services over data.
- Research funder policies increasingly mandate data stewardship of data produced by funded projects.
- Increasing need for data repositories and data stewardship.
  - Increasing volume presents a challenge.
  - Requirements for stewardship present a greater challenge.
- **Sustaining digital data infrastructure is a major issue for science policy!**
- Genuine concern that current funding models will prove inelastic and not meet the growing requirements – concern on the part of repositories and funders.
- **Important to demonstrate the value proposition of data repositories / data services.**
- **Sustainability is not just about whether something is funded, but how it is funded: what are the most effective and sustainable mechanisms for funding?**

OECD Global Science Forum Project: Business Models for Sustainable Data Repositories

- Relatively little work has been done on the economics and business models of data infrastructure.
  - Blue Ribbon Task Group Report on Sustainable Digital Preservation: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf
    - **Need to understand value proposition for communities.**
  - Sustaining Domain Repositories for Digital Data: A White Paper (ICPSR): http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf
    - **Need to understand how repositories are funded.**
- OECD Project builds on previous work of RDA-WDS Interest Group: 'Income Streams for Data Repositories': http://dx.doi.org/10.5281/zenodo.46693

OECD Global Science Forum Project:
Business Models for Sustainable Data Repositories

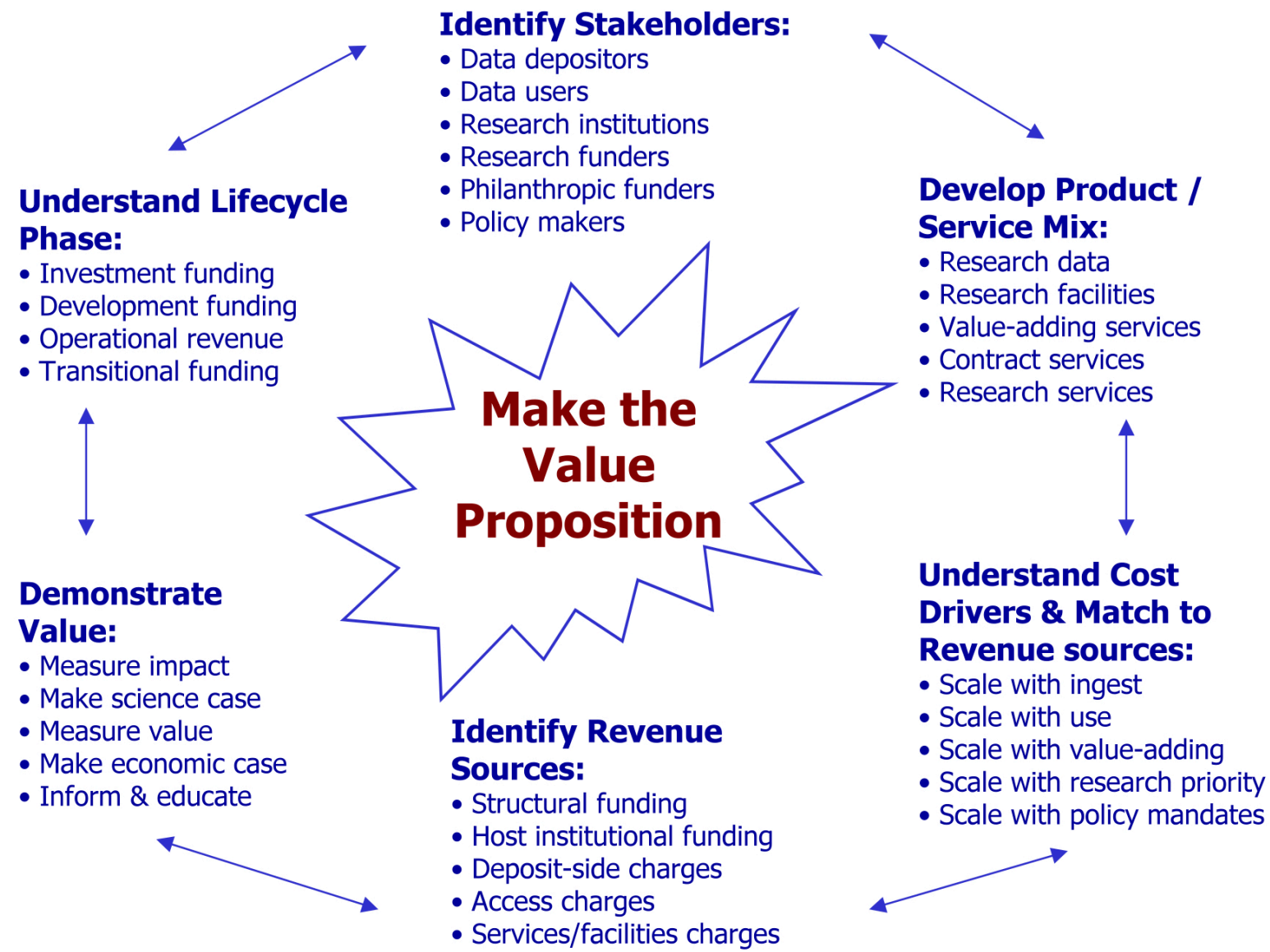
Central questions:

- How are data repositories currently funded?
- What innovative income streams are available to data repositories?
- What means of optimising costs are available?
- How do income streams match willingness/ability to pay of various stakeholders?
- How do income streams/willingness to pay fit together into a sustainable business model?

Deliverable:

- Survey of c.50 data repositories funding and business models.
- Two consultative workshops, SWOT and economic analysis.
- Report summarizing findings and containing policy recommendations for OECD member states to promote sustainable business models for data infrastructures.
- **Report to be released towards the end of 2017.**

OECD Global Science Forum Project: Business Models for Sustainable Data Repositories



Data is difficult: motivations and reward

- Open and FAIR data is essential for transparency and reproducibility; to take advantage of analysis at scale; to tackle major interdisciplinary challenges that require integration of data from many resources; has significant economic and other societal benefits…
- **But…**
- Research funders and research performing institutions will have to invest in data infrastructure.
- Essential to consider the cost of data stewardship and dissemination as part of the total cost of doing research.
- Data description, definitions and ontologies, data management require significant effort.
- Requires data skills, motivation and reward.
- Data should be integrated more with the process of scholarly communication and recognition of research contribution: **data citation** and journal availability policies; recognition for making available major datasets.
- **RPOs and research groups will increasingly build prestige on the basis of their data collections: research intensive institutions will be data intensive institutions.**

Incentives: Data Citation

If publications are the stars and planets of the scientific universe, data are the 'dark matter' – influential but largely unobserved in our mapping process

DC1 Data Citation Principles

FORCE11 The Future of Research Communications and e-Scholarship

CODATA Task Group on Data Citation Principles and Practices

Out of Cite, Out of Mind
http://bit.ly/Out_of_Cite_Report

Joint Declaration of Data Citation Principles:
https://www.force11.org/datacitation

Background and Developments:
http://bit.ly/data_citation_principles

International Series of Data Citation Workshops
http://bit.ly/data-citation-workshops

CODATA 2017

http://codata2017.gcras.ru/

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

CODATA 2017: Global Challenges and Data Driven Science

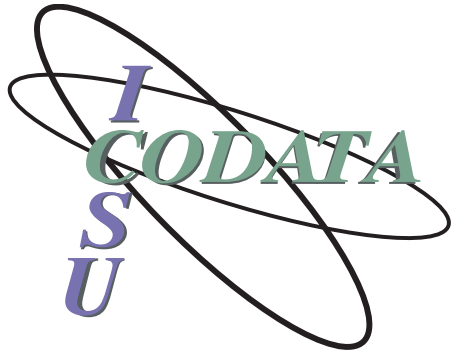
Major conference themes:

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 28 July 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!

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

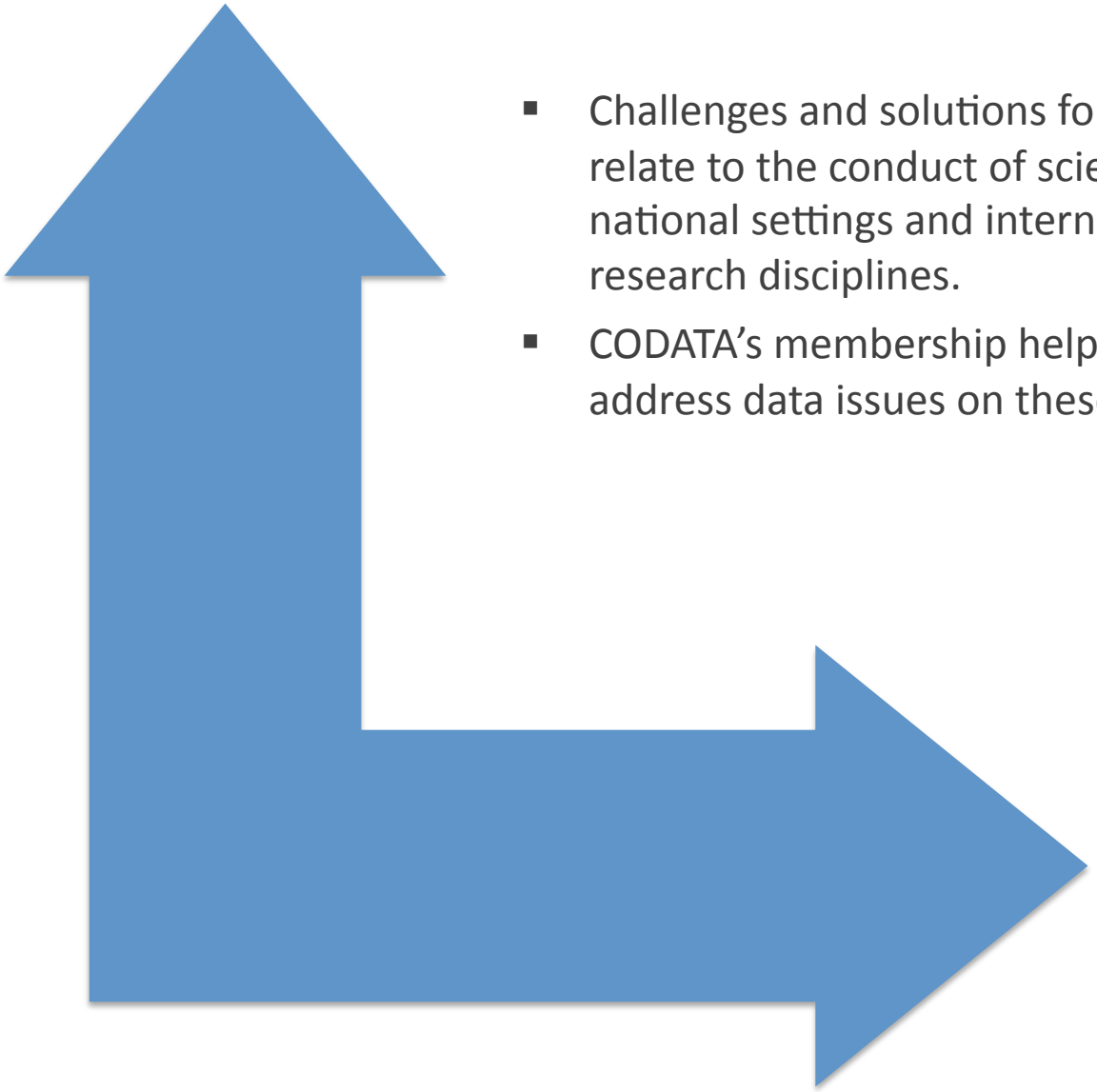
CODATA (ICSU Committee on Data for Science and Technology), 5 rue Auguste Vacquerie, 75016 Paris, FRANCE

Research Data: challenges and stakeholders

National Research Systems

CODATA National Members

National Academies of Science or Data Organisations



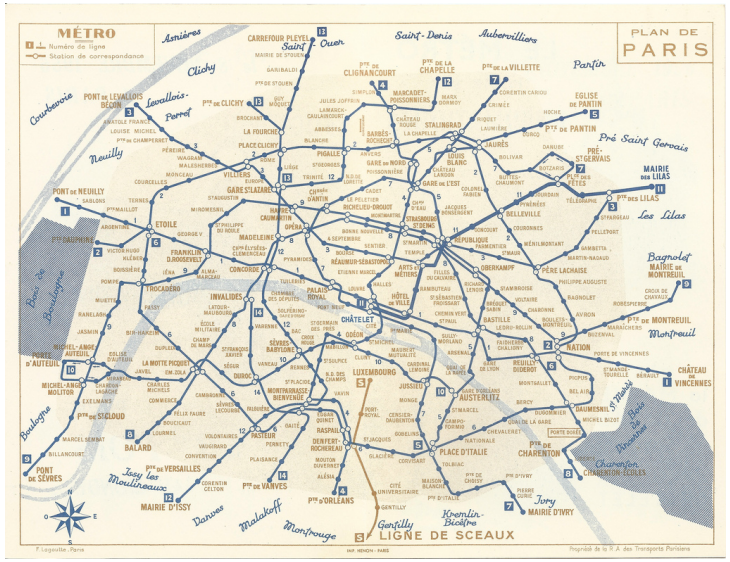
- Challenges and solutions for data issues relate to the conduct of science in national settings and international research disciplines.
- CODATA's membership helps us to address data issues on these two axes.

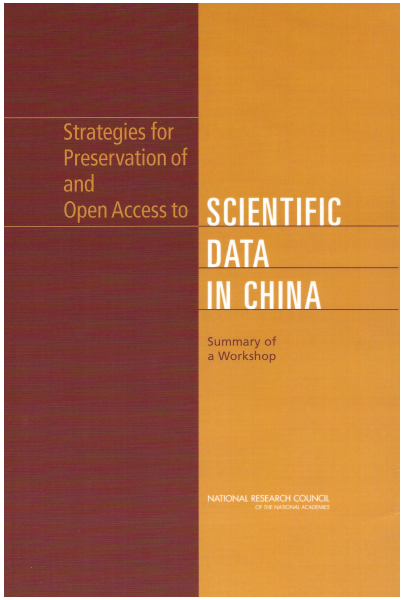
Scientific Disciplines

CODATA International Scientific Union Members

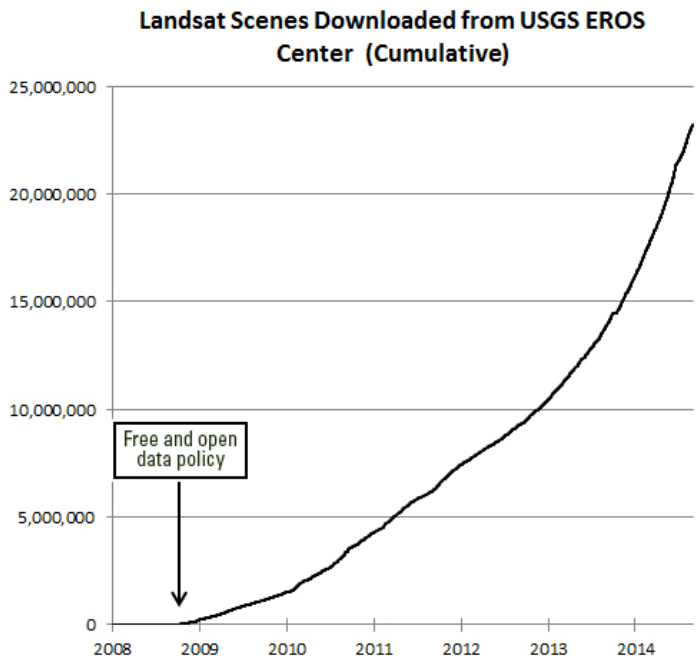
Role of CODATA National Committees

- **Join CODATA and form a National Committee**.
  - CODATA membership dues are aligned with GDP.
  - CODATA National Committees are composed of national stakeholders and data experts.
- **What are the benefits of having a CODATA National Committee?**
  - **Engage:** point of contact with CODATA;
  - **Influence:** contribute to CODATA strategy;
  - **Coordinate: forum by which national stakeholders may advance data agenda in step with international developments;**
  - **Collaborate:** propose Task Groups, host or participate in international workshop series, engage with Early Career Data Professionals Group;
  - **Partner:** undertake activities with other National Committees, bilaterally or in groups.



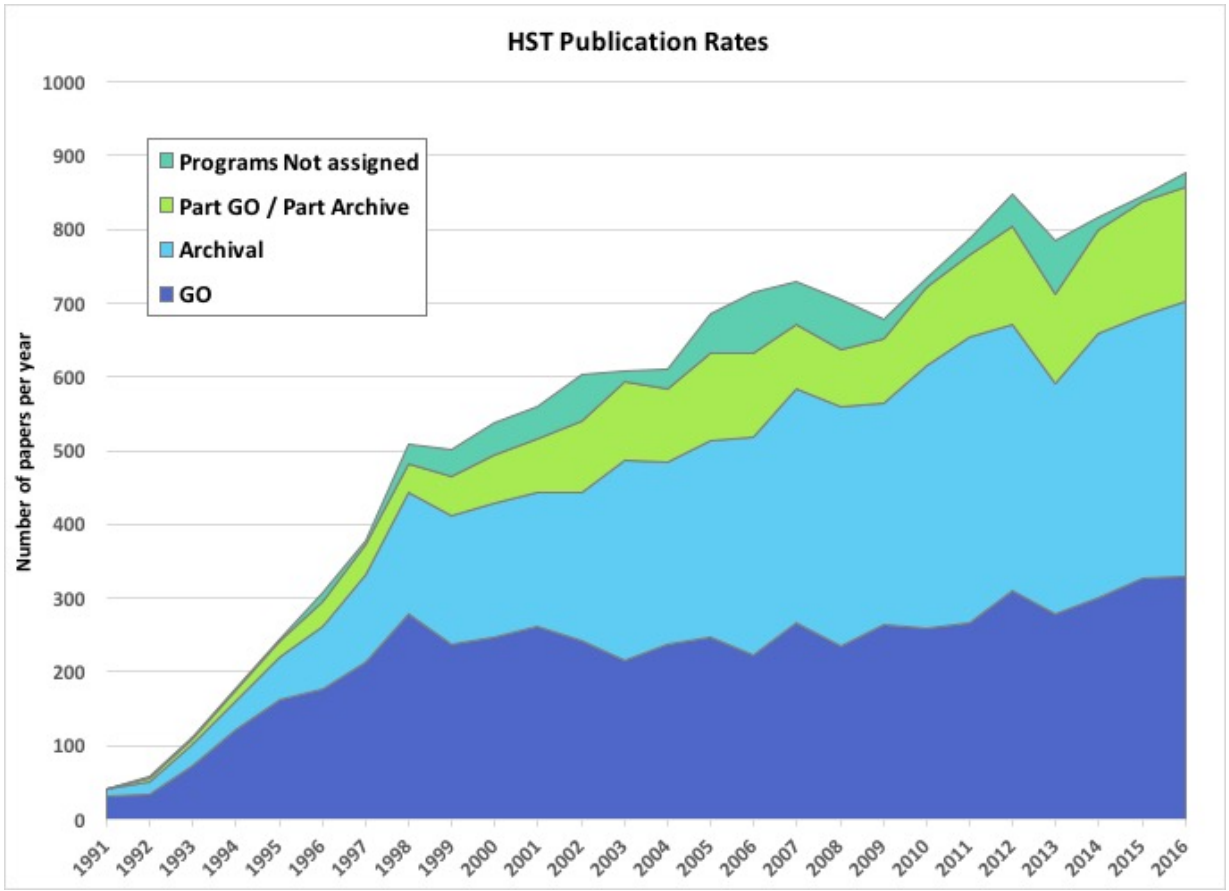


Economic Benefits of Data Sharing: LandSat



- **2006 Study** estimated the loss in case of a data gap as equivalent to US$935 M.
- **2011 Study** estimated benefits of landsat-sourced information for agriculture as US$858 M just for the state of Iowa.
- **2015 Study** estimated worldwide economic benefit of US$2.19 BN.
- Estimated benefit in US of US$1.8 BN.
- Valuing Geospatial Information: Using the Contingent Valuation Method to Estimate the Economic Benefits of Landsat Satellite Imagery: http://dx.doi.org/10.14358/PERS.81.8.647 **(Paywall… Irony…)**
- **Open data and open data infrastructure has a significant economic benefit.**

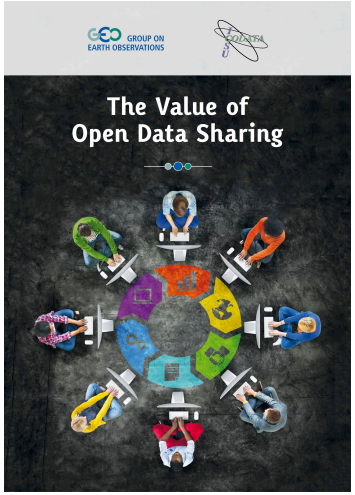
Reuse of Hubble Data for Different Purposes



Papers based upon reuse of archived observations now exceed those based on the use described in the original proposal: http://archive.stsci.edu/hst/bibliography/pubstat.html

CODATA Data Policy Activities

- New Data Policy Committee, chaired by Paul Uhlir, international expert in Data Policies and member of CODATA Executive Committee.
- Current Best Practice for Research Data Management Policies http://dx.doi.org/10.5281/zenodo.27872
- The Value of Open Data Sharing, report for GEO http://dx.doi.org/10.5281/zenodo.33830
- Legal Interoperability, Principles and Implementation Guidelines https://doi.org/10.5281/zenodo.162241
- FAIR Data
  - Simon Hodson is chairing the European Commission's Expert Group on FAIR Data: http://bit.ly/FAIR_Data_Expert_Group
- OECD Global Science Forum and CODATA Project on Business Models for Sustainable Data Repositories: http://www.codata.org/working-groups/oecd-gsf-sustainable-business-models



Findable Accessible Interoperable Reusable

OECD