CODATA Seminar
Academica Sinica, Taipei
27 October 2017

FAIR data and the roles and responsibilities of research institutions

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

CODATA Strategy: Mobilising the Data Revolution

- A global organisation, NFP, founded by the International Council of Science to address data issues.
- Member countries and participation in TGs, WGs and other initiatives comprises all continents.
- Executive Committee has wide representation, includes members from **Kenya** and **South Africa**.

Three priority areas essential to a coordinated international response to the data revolution.

- ➢ Promoting implementation of open data principles, policies and practices;
- ➢ Advancing the frontiers of data science and adaptation to scientific research;
- ➢ Building capacity by improving data skills and the functions of science systems needed to support open data (particularly in LMICs)

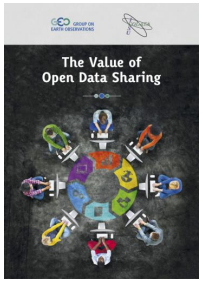
CODATA Prospectus:

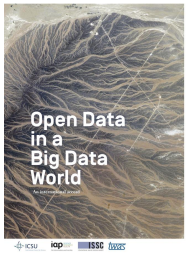
https://doi.org/10.5281/zenodo.165830

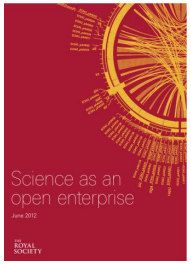
Principles, Policies and Practice

DC1 Data Citation Principles

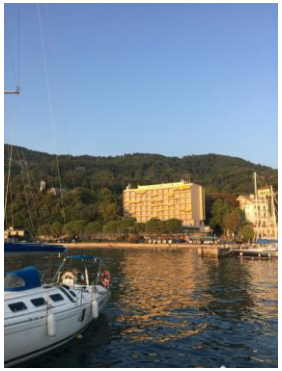
INTERNATIONAL COUNCIL FOR SCIENCE







Capacity Building

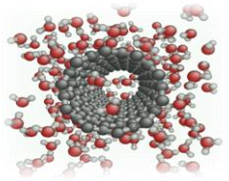


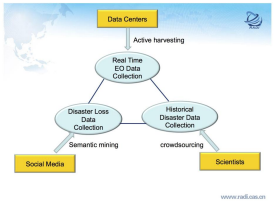
CHINESE ACADEMY OF SCIENCES

UNESCO

The Abdus Salam International Centre for Theoretical Physics

Frontiers of Data Science





ubiquity press open scholarship

Data Science Journal

CODATA 2017, Saint Petersburg 8-13 Oct 2017

ICSU WORLD DATA SYSTEM

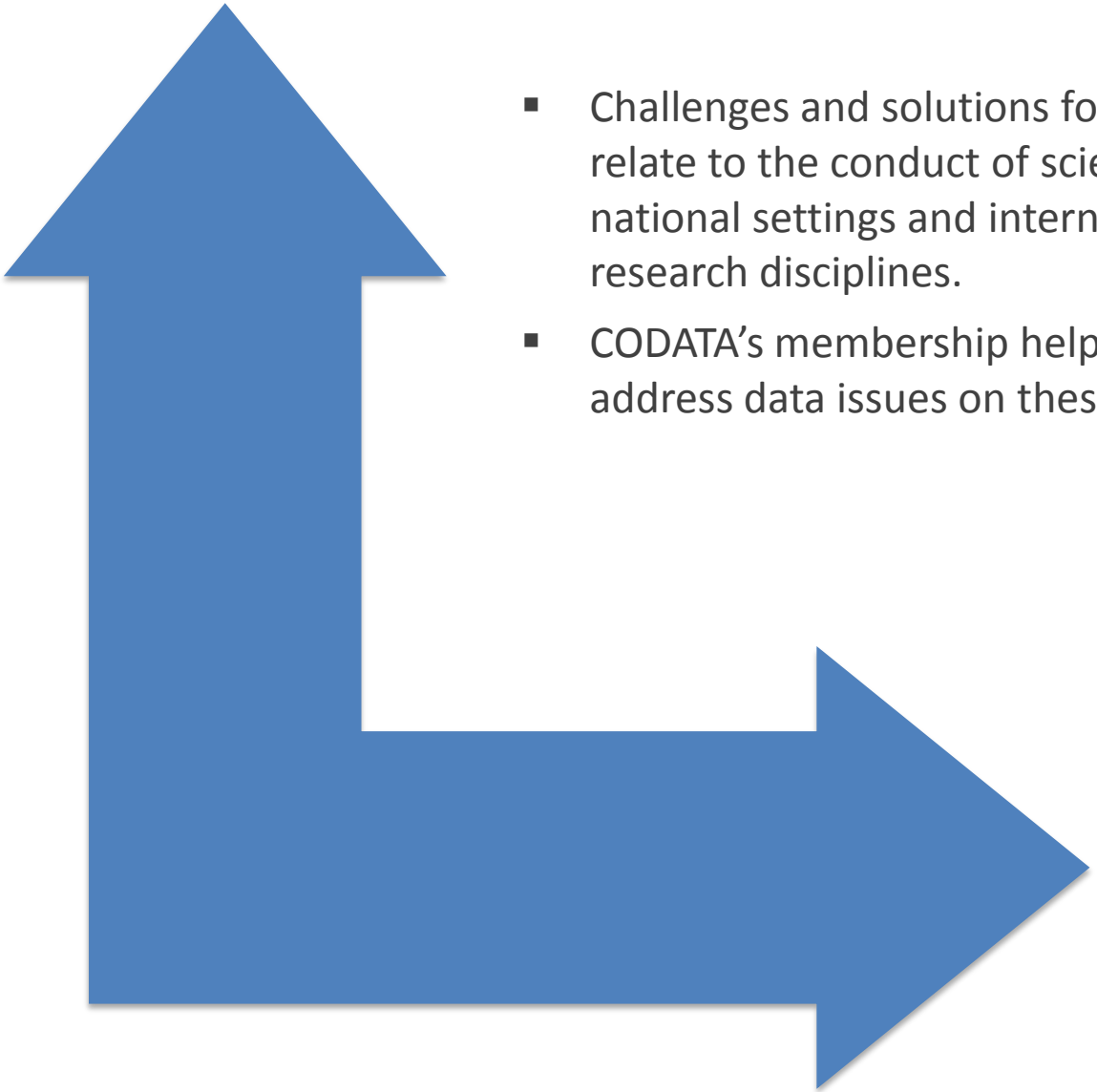
RDA RESEARCH DATA ALLIANCE

Research Data: challenges and stakeholders

National Research Systems

CODATA National Members

National Academies of Science or Data Organisations



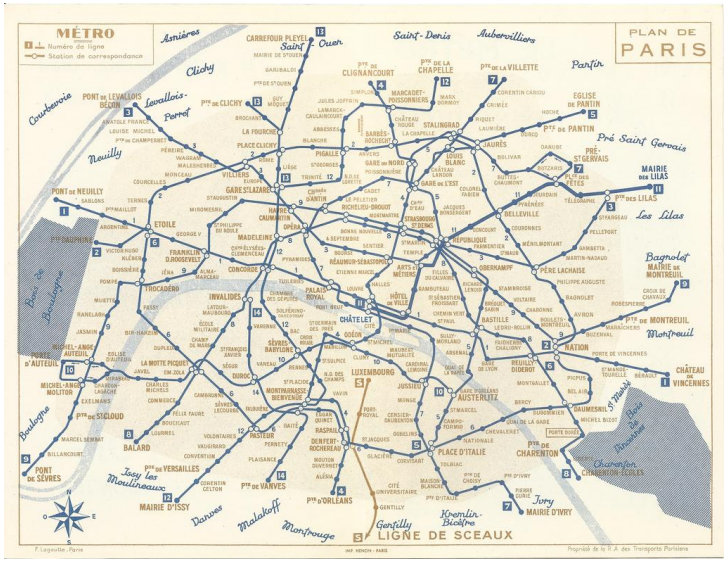
- Challenges and solutions for data issues relate to the conduct of science in national settings and international research disciplines.
- CODATA's membership helps us to address data issues on these two axes.

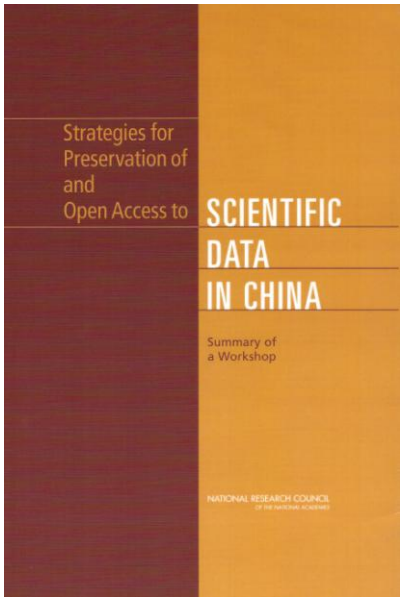
Scientific Disciplines

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.





Leverage

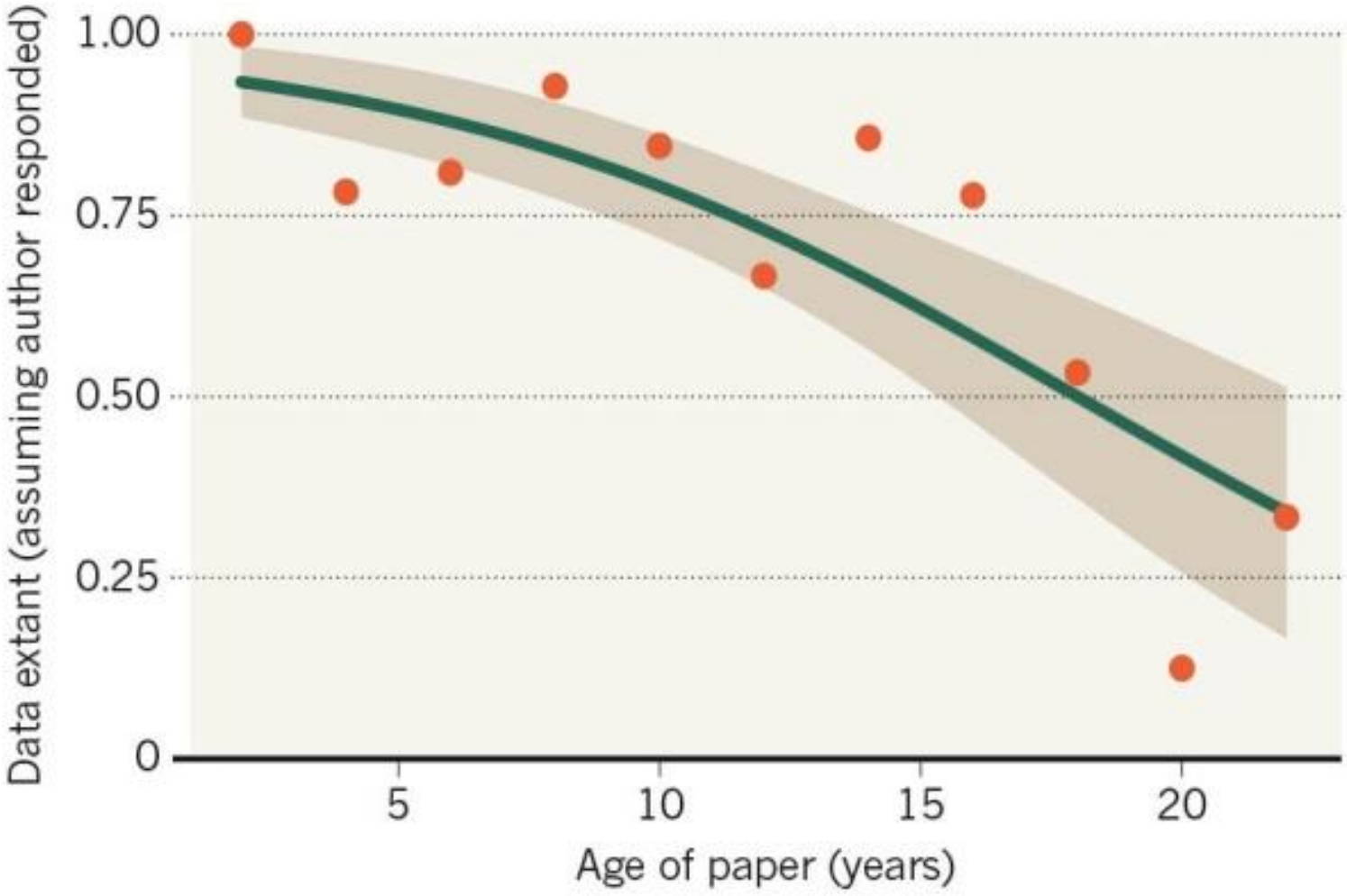
- **From September 2015 to September 2016, the annual income from membership fees of c.€205K leveraged further investment in activities to a total of over €1.9M: a leverage ratio of over 9.6:1**. This estimate includes external contributions to events, Task Groups and similar activities, sponsorship obtained as well as host and participant investment in events. As a specific example, in August 2016, the CODATA-RDA School of Research Data Science was held at ICTP in Trieste. CODATA's own investment in the event totals c.€10K in travel and student support. The event as a whole leveraged an additional c.€270,000 in support, comprising international and local travel and accommodation for experts and students as well as sponsorship and local expenses. **It should be noted that this estimate considerably undervalues the CODATA's leveraging power as it does not include any estimate for contributions in kind (e.g. co-chairs time).**

- **Next two years, concerted outreach to expand membership and to engage more with National Committees.**

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.

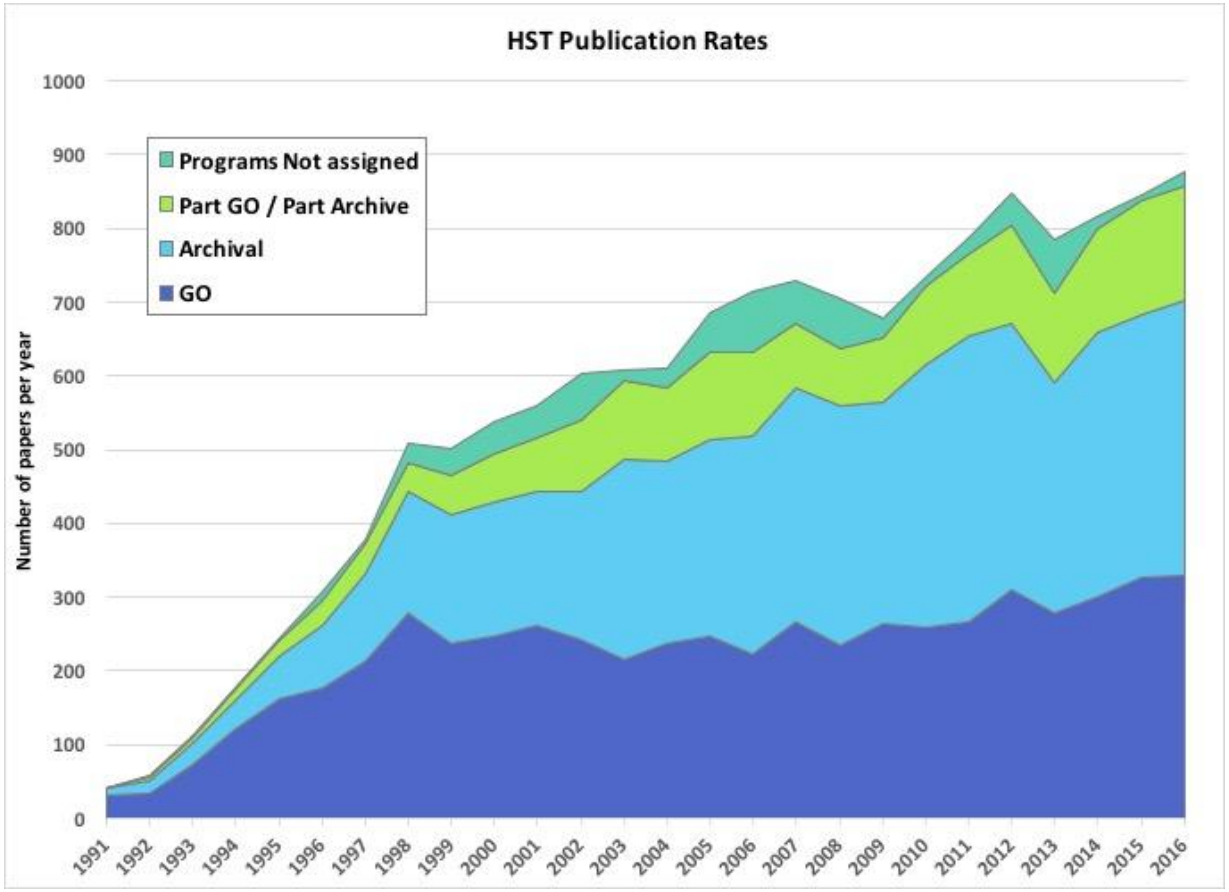
80% of ecology data irretrievable after 20 years (516 studies)



Vines TH *et al*. (2013) Current Biology DOI: 10.1016/j.cub.2013.11.014

Slide credit: Dryad

Reuse of Hubble Data for Different Purposes

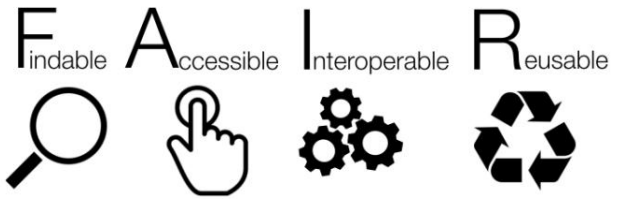


Papers based upon reuse of archived observations now exceed those based on the use described in the original proposal: http://archive.stsci.edu/hst/bibliography/pubstat.html

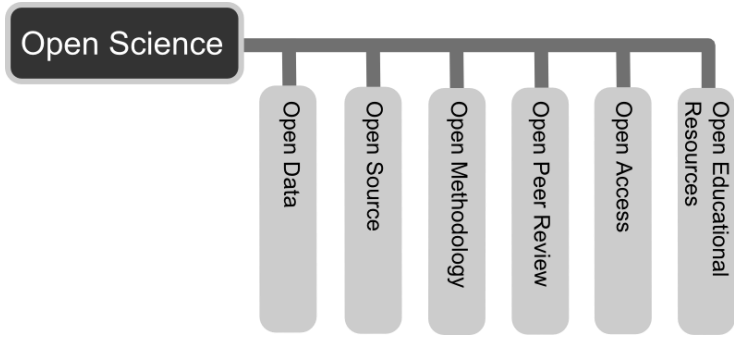
Policy Push for Open Research Data

- 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

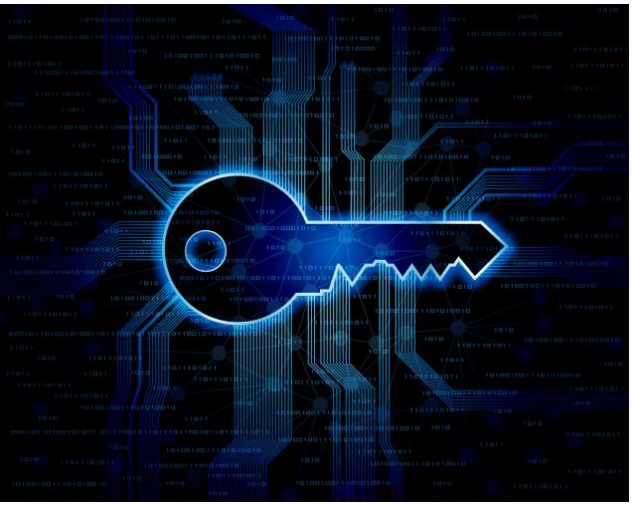
Open Science



- What is Open Science:
  - Open access to research literature.
  - **Data that is as Open as possible, as closed as necessary.**
  - **FAIR Data (Findable, Accessible, Interoperable, Reusable).**
  - A shop window and repository of all research outputs.
  - A culture and methodology of open discussion and enquiry (including methodology, lab notebooks, pre-prints)
- Research data is evidence: it is fundamental to the validity and reproducibility of science.
- Those research disciplines that have leapt forward in the past 15-20 years are those that have shared and analysed data at scale: genomics, astronomy, disciplines using remote sensing data etc.
- **LAC research institutions have an opportunity to build their reputation around research specialisation: and this requires data specialisation and FAIR data collections.**

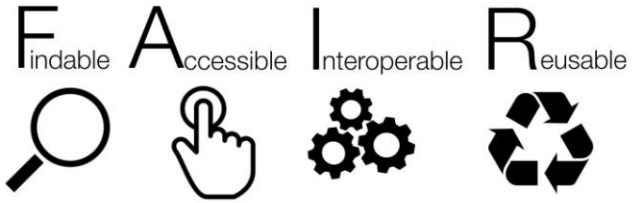


Boundaries of Open



- 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!
- **A great deal of data is not affected by these issues and can and should be open.**
- **There is a need to evolve policies, practices and ethics around closed, shared, and open data.**

FAIR Data



- **FAIR Data:** increasingly widely adopted as summary of attributes that increase value of research data.
- Chair of European Commission Expert Group on FAIR Data 'Making FAIR Data a Reality': http://bit.ly/FAIRdata-EG

- The value of data lies in reuse. What are the attributes that make data reusable?
  - **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.
- FAIR is augmented by the principle that data should be '**as open as possible, as closed as necessary'**, or 'open by default'.

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)

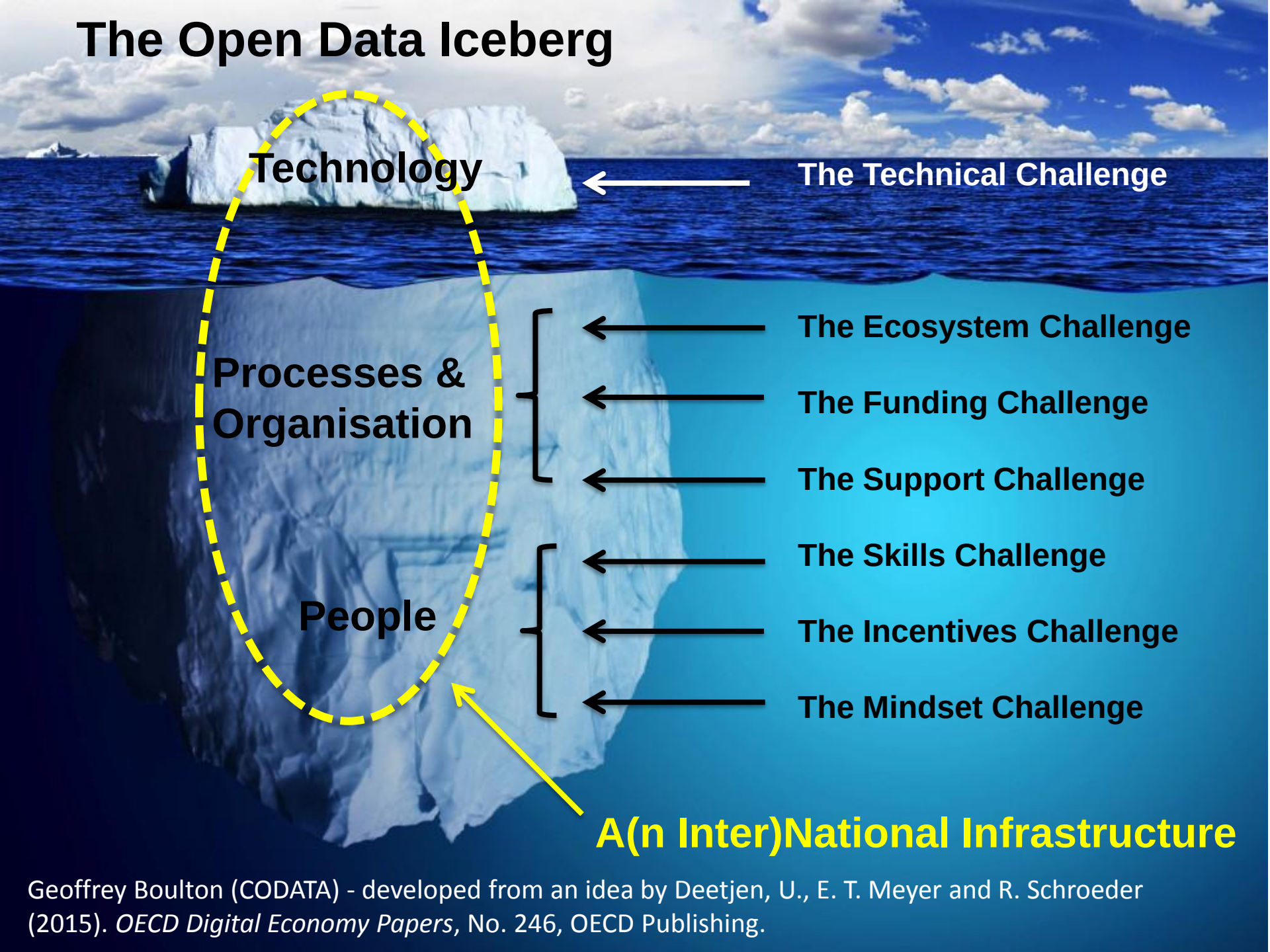
FAIR Guiding Principles (2)

- **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 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, FAIR data
- **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.

Framework for National and Institutional Data Strategies

- National / Institutional Open and FAIR Data Strategy.
- Open data **policies and guidance** at national and institutional level.
- Clarify the **boundaries of open** (particularly privacy, IPR).
- Mechanisms (infrastructure and policy) to ensure **concurrent publication of data as research output**.
- Data 'publication' and citations of data included in **assessment of research contribution**.
- Promotion of **data skills** (researchers and data stewards).
- Development **of institutional infrastructure** for research collaboration and data stewardship/RDM.
- **Collaborative infrastructures** for certain research disciplines, nationally, regionally to pool expertise and lower costs.

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.**

Göttingen-CODATA Symposium 18-20 March 2018

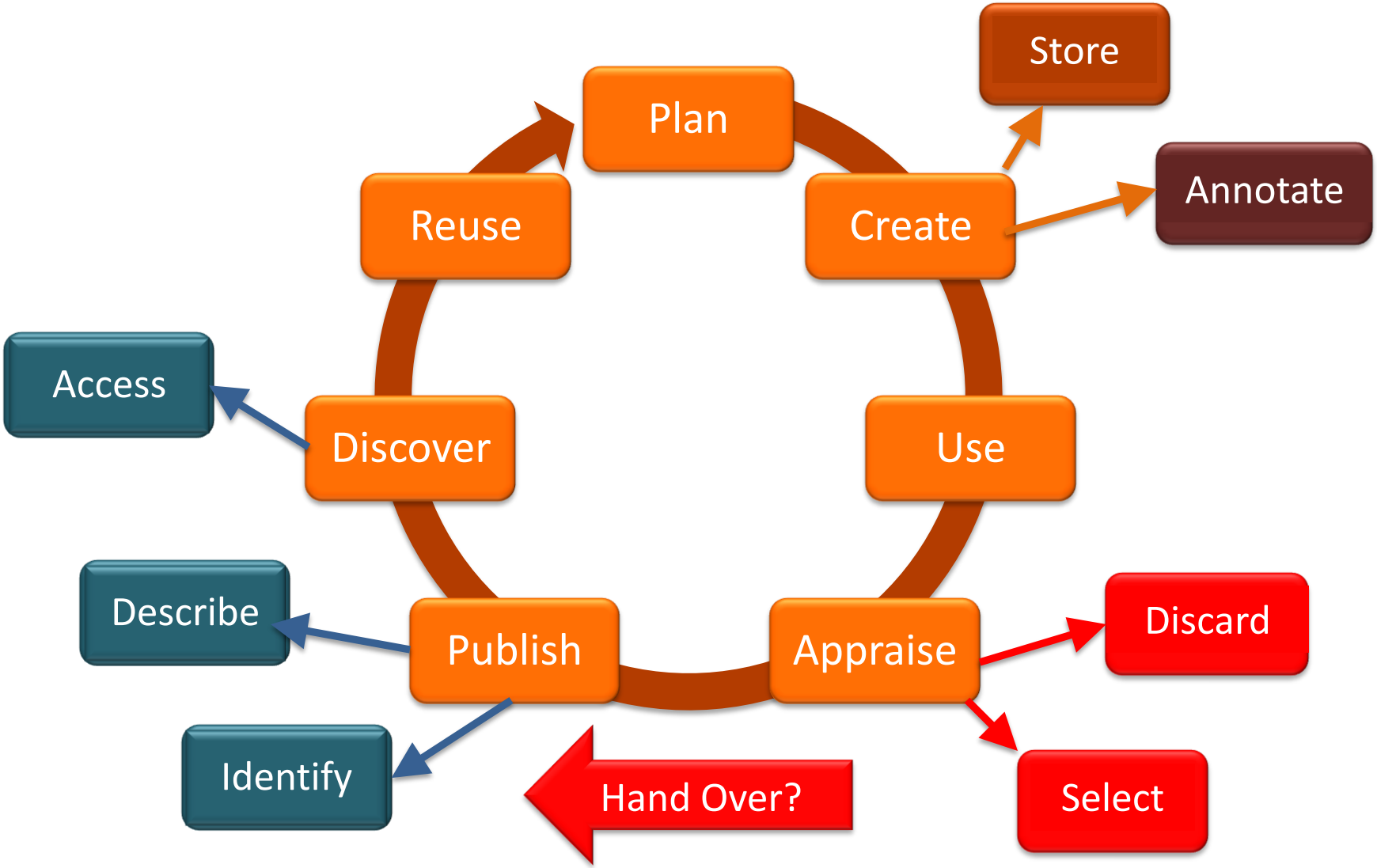
- **The critical role of university RDM infrastructure in transforming data to knowledge**
- http://conference.codata.org/2018-Goettingen-RDM/
- An opportunity to share experiences, research and insights in the development implementation of RDM services in research institutions.
- Special collection of Data Science Journal.
- Themes: services and solutions; strategy; measuring success; skills and support; sustainability; shared services and outsourcing / consortiums; service level, trust and FAIR; champions and engaging with researchers.
- Announcement of call for papers in the next week or so.
- Deadline for abstracts, 15 December.

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

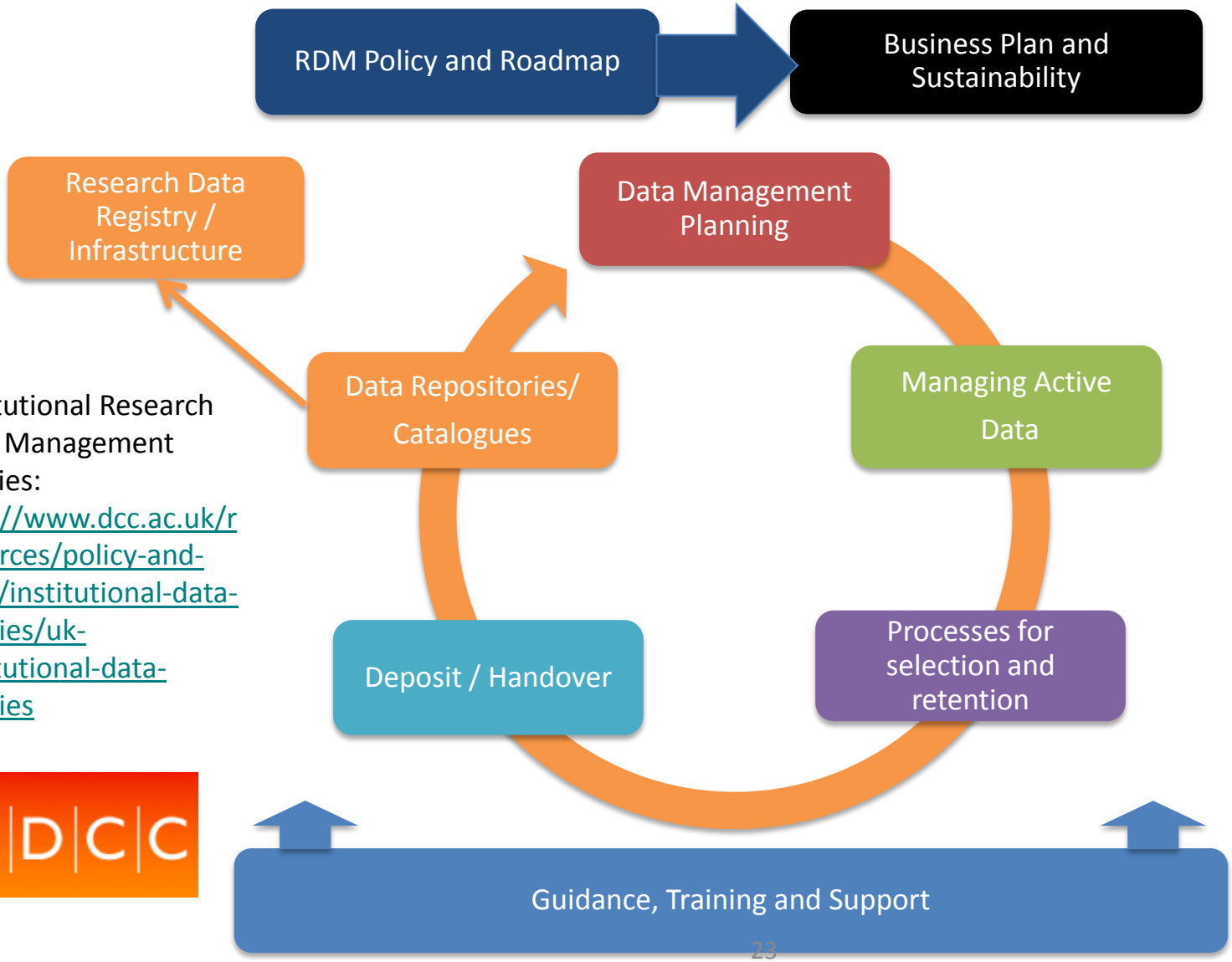
Where should research data go?

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)

Supporting the Research Data Lifecycle



Components of RDM support services



Institutional Research Data Management Policies: http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies/uk-institutional-data-policies

Barriers to Data Sharing

Researchers concerns:

- Concern that data may be misused or misunderstood.
- Concern that will lose scientific edge if sharing before fully exploited.
- Desire to retain control of a professional asset.
- Culture in particular research disciplines; availability of infrastructure.
- **Concern that will not be credited.**
- **Lack of career rewards for data publication.**
- **Fundamentally, researchers are reluctant to expend effort sharing data because they do not feel that data is adequately exposed or credited.**

Nature special issue on data sharing: http://www.nature.com/news/specials/datasharing/index.html

- See ODE report, using Parse.Insight findings: http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

Reuse of data and research parasites

'A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as "research parasites."'

EDITORIAL 'Data Sharing', Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D. N Engl J Med 2016; 374:276-277January 21, 2016 DOI: 10.1056/NEJMe1516564

Slide credit: Danny Kingsley

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

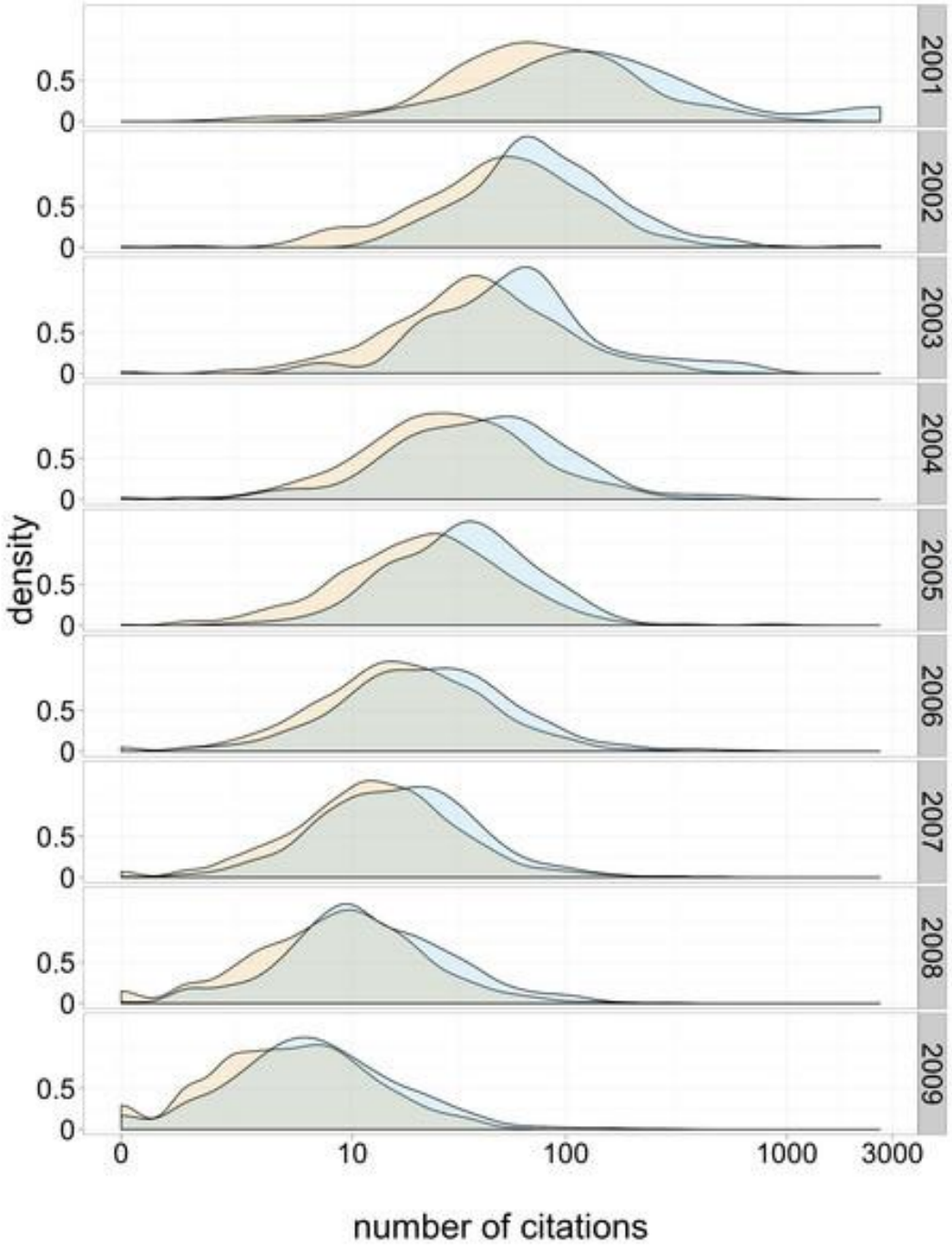
CODATA Task Group on Data Citation Principles and Practices

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



Citation advantage of having archived Gene Expression Omnibus data

Examined **10,555** studies that created gene expression microarray data, comparing those that made data available and those that didn't.

Studies that made data available in a public repository received 9% more citations than similar studies for which the data was not made available.
Increased citation of 30% for those published 2004-5.

Piwowar and Vision (2013), PeerJ DOI:10.7717/peerj.175

CODATA-RDA School of Research Data Science

- Contemporary research – particularly when addressing the most significant, transdisciplinary research challenges – increasingly depends on a range of skills relating to data. **These skills include the principles and practice of Open Science and research data management and curation, the development of a range of data platforms and infrastructures, the techniques of large scale analysis, statistics, visualisation and modelling techniques, software development and data annotation.** The ensemble of these skills, relating to data in research, can usefully be called 'Research Data Science'.

software carpentry

DATA CARPENTRY
MAKING DATA SCIENCE MORE EFFICIENT

DCC

Foundational Curriculum

Seven components: open science, data management and curation; software carpentry; data carpentry; data infrastructures; statistics and machine learning; visualisation.

Builds on much existing courses to create something more than the sum of its parts:

- **Open Science** – reflection on ethos and requirements of sharing/openness
- **Open Research Data** – Basics of data management, DMPs, RDM life-cycle, data publishing, metadata and annotation
- **Author Carpentry** – Improving research efficiency with command line and OS tools.
- **Software Carpentry** – Introduction the Unix shell and Git (sharing software and data)
- **Data Carpentry** – Introduction to programming in R, and to SQL databases
- **Visualisation** – Tools, Critical Analysis of Visualisation
- **Analysis** – Statistics and Machine Learning (clustering, supervised and unsupervised learning)
- **Computational Infrastructures** – Introduction to cloud computing, launching a Virtual Machine on an IaaS cloud

Building international network of short courses http://bit.ly/first_data_school_trieste

Programme and materials: **http://bit.ly/School_of_Research_Data_Science-Programme** ; **http://bit.ly/first_data_school_materials**

#DataTrieste

CODATA-RDA Data Science Training Initiative

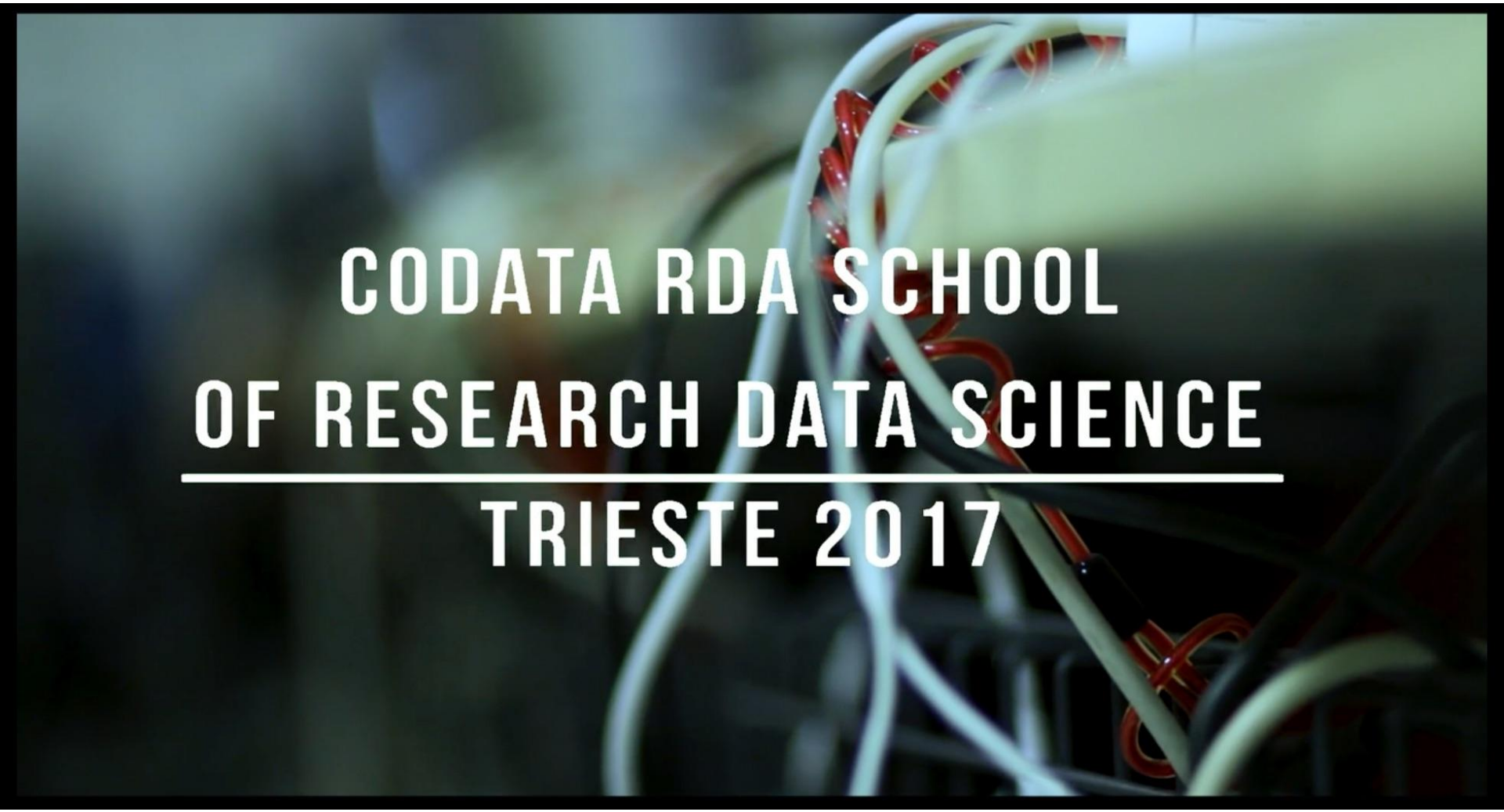
- Annual foundational school hosted at ICTP, Trieste (with the objective to build a network of partners, train-the-trainers).
- Advanced workshops, ICTP, Trieste, following the foundational school.
- National or regional schools, organised with local partners.
- Planning at least two pilot schools as part of the African Open Science Platform project.

- Next #DataTrieste Summer School, 6-17 August 2018.
- Next #DataTrieste Advanced Workshops 20-24 August 2018.
- Next regional foundational school 'CODATA-RDA School of Research Data Science', São Paulo, 4-15 December 2017: http://www.ictp-saifr.org/?page_id=15270 - **deadline 22 September**

The Abdus Salam International Centre for Theoretical Physics

twas

DataTrieste Film on Vimeo
https://vimeo.com/232209813



The Abdus Salam International Centre for Theoretical Physics

twas

INTERNATIONAL DATA WEEK
IDW 2018

Gaborone, Botswana: 22–26 October 2018

ICSU CODATA

ICSU WORLD DATA SYSTEM

RDA RESEARCH DATA ALLIANCE

UNIVERSITY of BOTSWANA

Digital Frontiers of Global Science

Frontier issues for research in a global and digital age.

Applications, progress and challenges of data intensive research.

Data infrastructure and enabling practices for international and collaborative research.

Botswana, Africa and the World!

Stable, safe, modern and exciting country

INTERNATIONAL COUNCIL FOR SCIENCE

ICSU CODATA

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