Open Science Conference 2018
Berlin
14 March 2018

Making FAIR data a reality… and the challenges of interoperability and reusability

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

INTERNATIONAL
COUNCIL
FOR SCIENCE

CODATA Prospectus:

https://doi.org/10.5281/zenodo.1167846

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

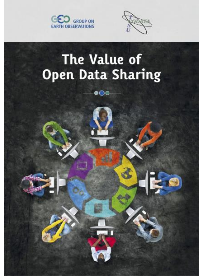
May 2014

DC1
Data Citation Principles

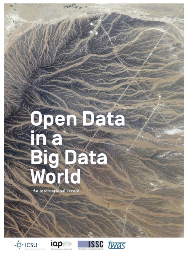
INTERNATIONAL COUNCIL FOR SCIENCE

OECD

GEO


GEO GROUP ON EARTH OBSERVATIONS

The Value of Open Data Sharing


Open Data in a Big Data World

An international accord

ICSU iap ISSC twas

OECD publishing

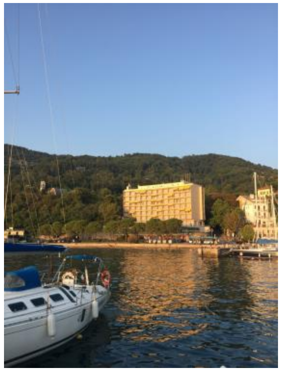
BUSINESS MODELS FOR SUSTAINABLE RESEARCH DATA REPOSITORIES

OECD SCIENCE, TECHNOLOGY AND INNOVATION POLICY PAPERS

December 2017 No. 47

OECD

Capacity Building



CODATA

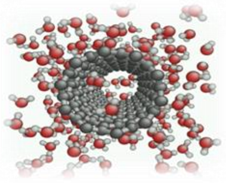
中國科学院
CHINESE ACADEMY OF SCIENCES

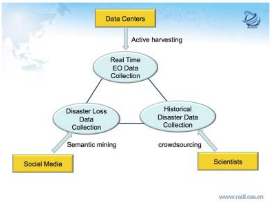
UNESCO

GO FAIR

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
futurearth
IRDR
北京国际会议中心

]u[ubiquity press
open scholarship

ISSN 1234-1234 (online)

Volume 1 (2015)

DATA SCIENCE JOURNAL

Editor in Chief: Sarah Callaghan

]u[

Data Science Journal

International Conference
CODATA 2017

"GLOBAL CHALLENGES
and DATA-DRIVEN SCIENCES"

CODATA 2017, Saint Petersburg 8-13 Oct 2017

INTERNATIONAL DATA WEEK
IDW 2018

Gaborone, Botswana: 22–26 October 2018

CODATA ICSU WORLD DATA SYSTEM RDA RESEARCH DATA ALLIANCE UNIVERSITY OF BOTSWANA

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.

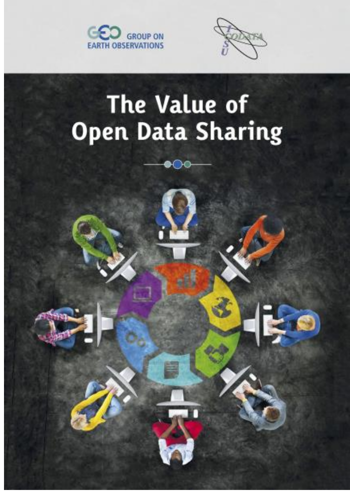
Science as an open enterprise
June 2012
THE ROYAL SOCIETY
A WORLD THAT COUNTS
MOBILISING THE DATA REVOLUTION FOR SUSTAINABLE DEVELOPMENT
GODAN
ODI
How can we improve agriculture, food and nutrition with open data?
UK20150528
Open Data in a Big Data World
An international accord
ICSU
iap
ISSC
twas

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

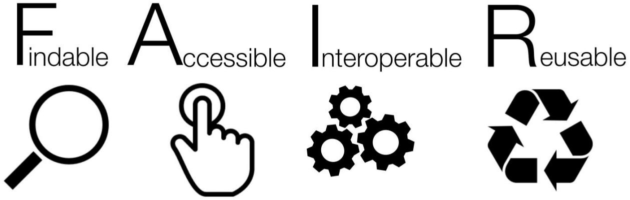
CODATA Data Policy Activities

- Current Best Practice for Research Data Management Policies http://dx.doi.org/10.5281/zenodo.27872
- Science International Accord on Open Data in a Big Data World: http://www.science-international.org/
- 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
- **Data Policy Committee, chaired by Paul Uhlir, international expert in Data Policies and member of CODATA Executive Committee.**



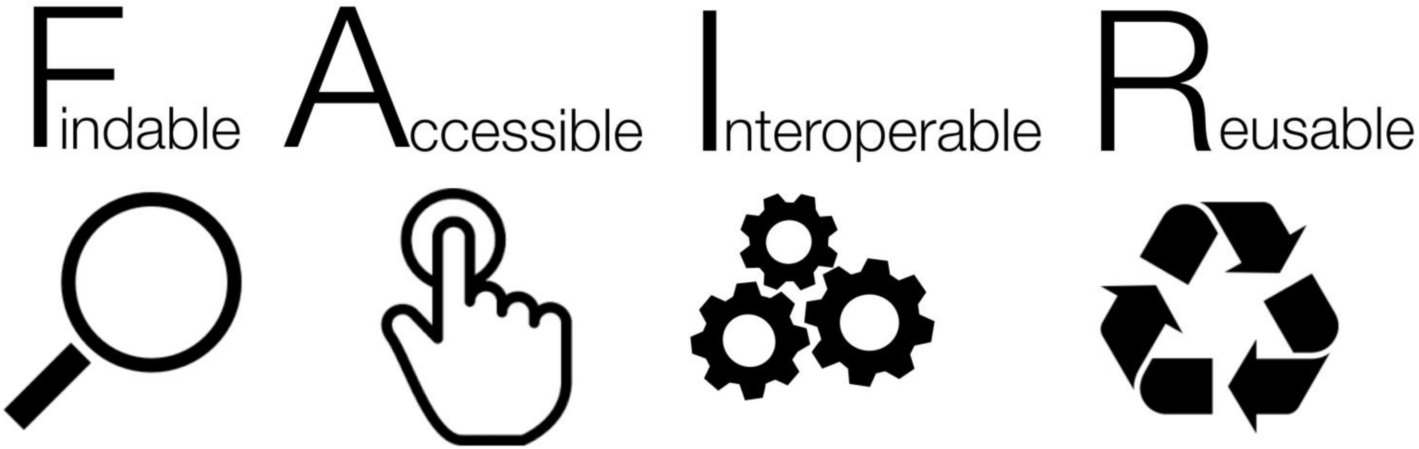
GEO GROUP ON EARTH OBSERVATIONS

The Value of Open Data Sharing



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.



Attributes that give value to research data

- Builds on previous definitions…
- OECD Statement of Principles and Guidelines for Access Research Data: include a number of principles including accessibility, interoperability, quality, legal transparency, sustainability…
- Royal Society, 2012, *Science as an Open Enterprise*, Intelligent Openness: **accessible, intelligible, assessable, usable**.
- G8 Science Ministers' Statement, 2013, 'Open scientific research data should be easily **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

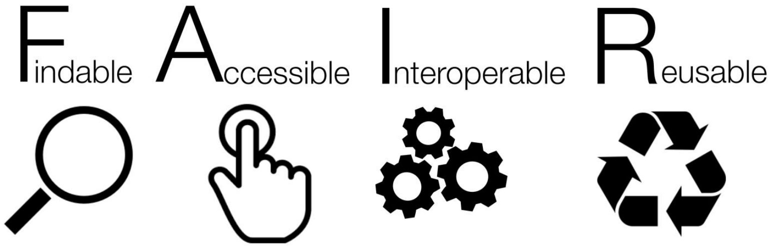
European Commission Expert Group on FAIR Data

Core Deliverables

1. To develop **recommendations** on what needs to be done to turn each component of the FAIR data principles into reality
2. To propose **indicators** to measure progress on each of the FAIR components
3. Actively **support the creation of the FAIR Data Action Plan**, by proposing a list of concrete actions as part of its Final Report.
4. Support Commission in **presentation of FAIR Data Action Plan** in Autumn 2018.

Related Deliverables

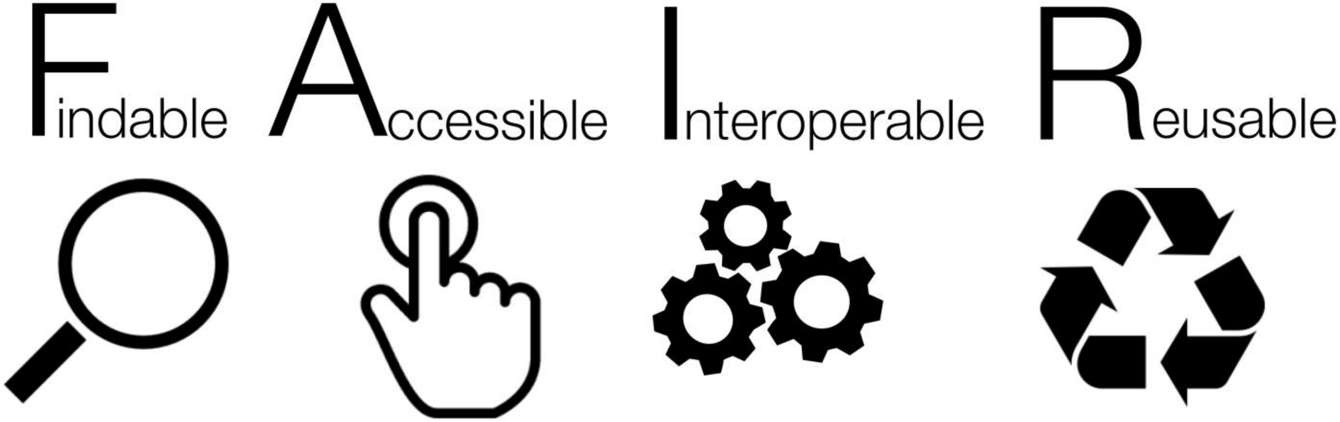
1. To contribute to the **evaluation of the Horizon 2020 Data Management Plan (DMP) template** and development of associated sector / discipline-specific guidance
2. To provide input on the issue of **costing and financing data management activities**



European Commission Expert Group on FAIR Data

Report Structure

1. Concepts: Why FAIR?
2. Creating a culture of FAIR data
3. Making FAIR data a reality: technical perspective
4. Skills and capacities for FAIR data
5. Measuring Change
6. **Facilitating Change: a FAIR Data Action Plan**



FAIR Data – Additional Concepts

- **As Open as Possible, as Closed as Necessary**
- **FAIR-TLC:** Traceability, Licensing, Connectedness (https://doi.org/10.5281/zenodo.203295)
- **FREE-FAIRER:** Findable, **Rapidly Available**, **Ethical**, **Equitable**; **Forever**, Accessible, Interoperable, **Reliable**.
- Timely release.
- Long term preservation in a trusted (sustainable) digital repository.
- Quality markers from community and repository.
- Responsibilities of users.
- …

- **Action Plan retain FAIR, retain the set of priorities it lays out, while discussing how FAIR can be achieved in a broader ecosystem.**

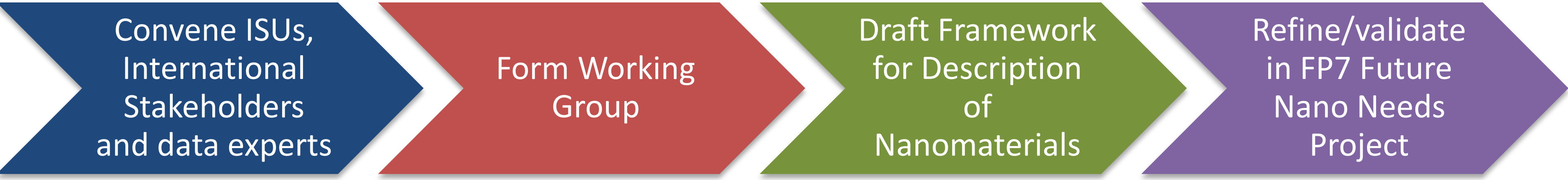
Open Science and FAIR Data: Implications and Directions

- Clarify the **boundaries of Open** for research data.
- Clarify and propagate **criteria and guidelines** for appraisal and selection.
- Recognise that data stewardship must be regarded as part of the total cost of doing research.
    - 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.
- Address the **skills** requirements for data scientists, data stewards, data liaisons and researchers themselves.
- **Refine and improve understanding of FAIR data, particularly I and R…**
- **Work with and across disciplines on standards and vocabularies, to help address I and R…**

Interoperability and Reusability

- The two most challenging areas of FAIR…
- Major challenges to clarify and unpack **interoperable** and **reusable**.
- **Need to build on knowledge from archival community: what makes a resource usable?**
- **OAIS Reference Model:**
  - ‘**Preservation Description Information** is divided into five types of preserving information called Provenance, Context, Reference, Fixity and Access Rights’
  - **Independently Understandable:** A characteristic of information that is sufficiently complete to allow it to be interpreted, understood and used by the Designated Community without having to resort to special resources not widely available, including named individuals.
  - **Designated Community**: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time.

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

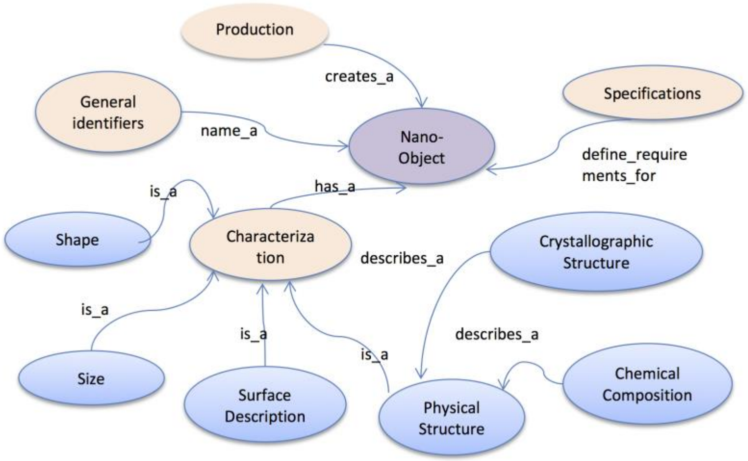


Figure 4. Information categories for describing an individual nano-object

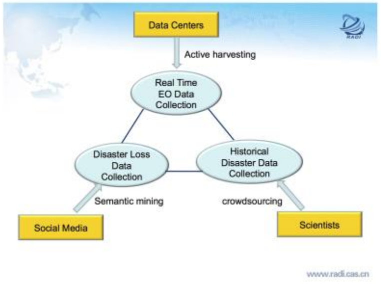
CODATA WG on the Description of Nanomaterials:
http://www.codata.org/nanomaterials

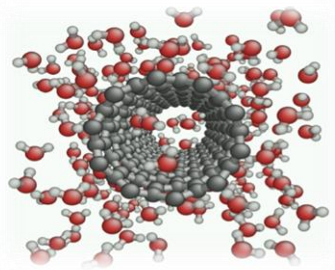
Uniform Description System v.02, May 2016:
http://dx.doi.org/10.5281/zenodo.56720

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

Interdisciplinary Research and Use Beyond the Designated Community

- Major interdisciplinary 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.





ICSU-CODATA Commission on Data Standards for Science

- **‘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
- Representatives of International Scientific Unions: IUCr for CIF; IUPAC for chemical terminologies; IUGS for GeoSciML; etc.
- Representatives of Standards Organisations: e.g. Darwin Core, for biology, biodiversity; DDI for social science surveys; OGC for geospatial data; W3C for the web.
- Report and Position Paper: https://doi.org/10.5281/zenodo.1193642
- Next Steps:
  - Directory of activities involving international scientific unions.
  - Maturity model for vocabularies and standards.
  - **Case studies of applications of vocabularies and standards for transdisciplinary research.**

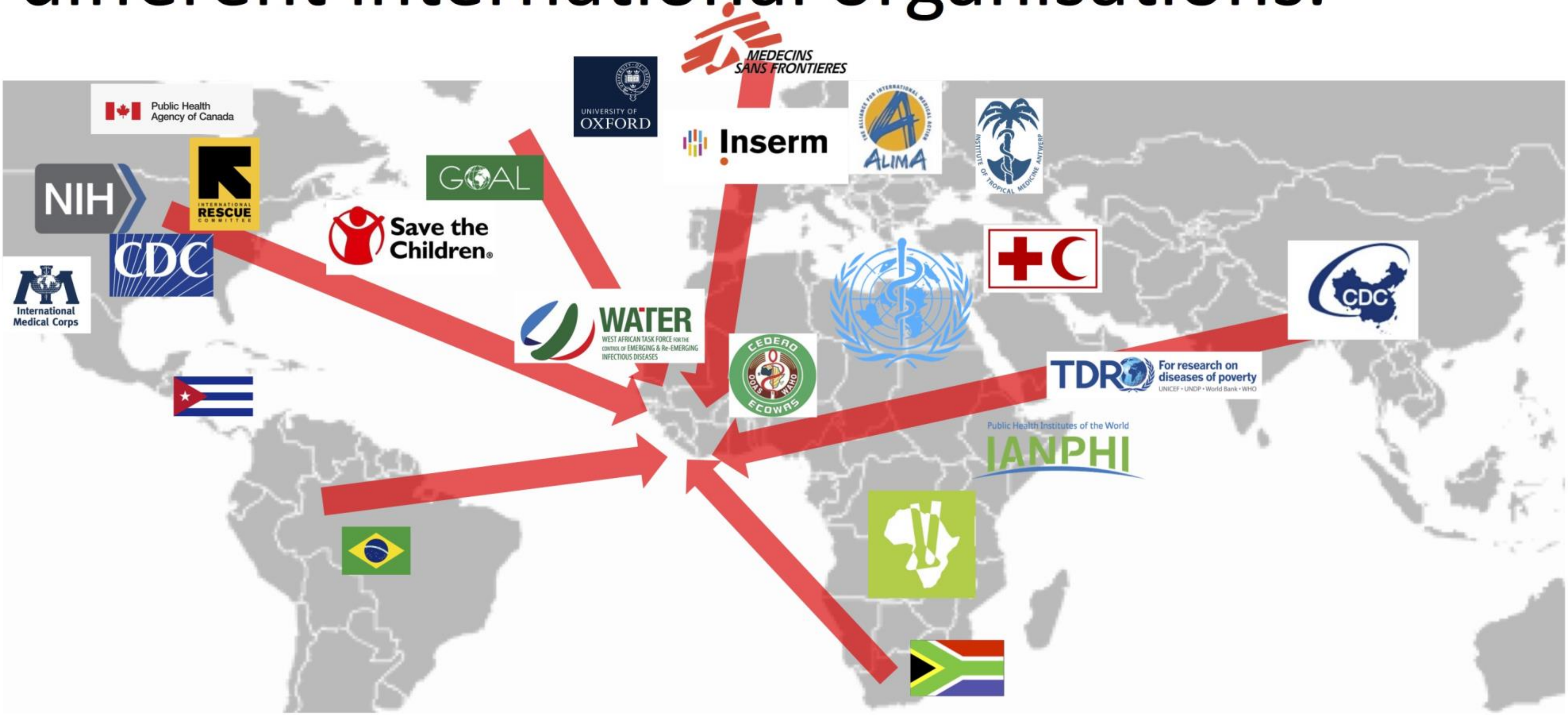
Initiative for Data Interoperability and Integration

- **Larger follow-up workshop 13-15 November, Royal Society, London.**
- Identified three-*four* pilots to explore interoperability and reuse in interdisciplinary research (a fourth in consideration):
  1. Understanding and responding to infectious disease outbreaks, particularly in crisis situations;
  2. Disaster risk research, particularly in relation to Sendai reporting;
  3. Research into resilient cities;
  4. Agricultural research particularly in developing countries.
- Vision of a decadal initiative to advance science through integration of data and information.

Initiative for Data Interoperability and Integration

- **Strand 1.** This **addresses important application domains** (infectious disease, resilient cities, and disaster risk, with a strong probability of adding agriculture): They have been chosen as major issues where relevant data exists and is accessible, where data integration is a tractable objective, and where there are existing communities of practice that are willing to collaborate.
    - Three stages: 1) a pilot project; 2) a full project; and, 3) a stage of high level integration between the domains
- **Strand 2.** Aims to **engage with disciplines of science** through the international unions, associations and other disciplinary bodies, seeks to evaluate the extent to which different disciplines address their data needs and opportunities, and will work to provide support to those disciplines that have not yet done this but wish to do so.
    - Drawing generic lessons on the actions and support required to promote interoperability and data integration.

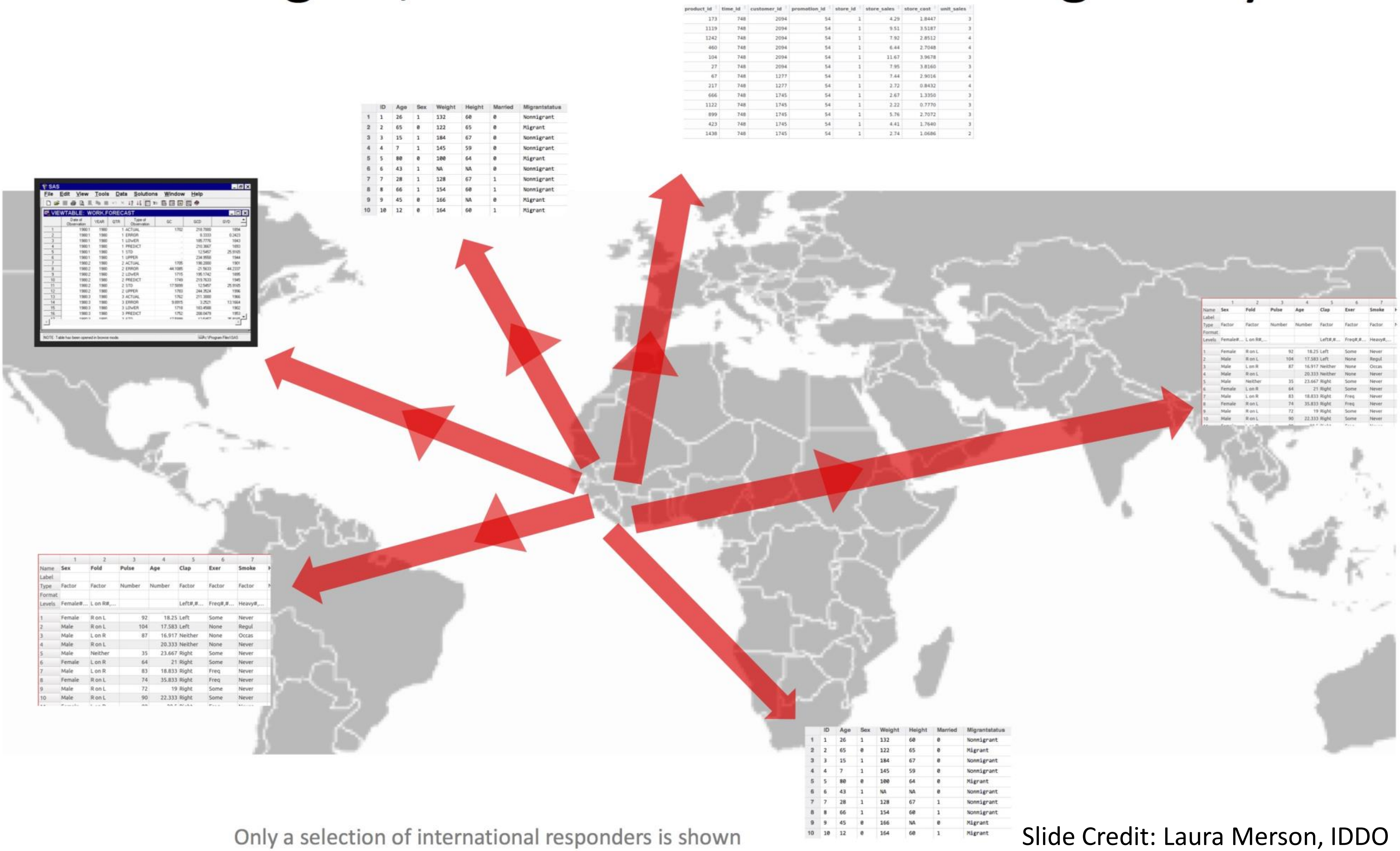
The government-led response to the West African Ebola outbreak included many different international organisations.


MEDECINS SANS FRONTIERES
UNIVERSITY OF OXFORD
Public Health Agency of Canada
Inserm
ALIMA
THE ALLIANCE FOR INTERNATIONAL MEDICAL ACTION
INSTITUTE OF TROPICAL MEDICINE ANTWERP
GOAL
NIH
INTERNATIONAL RESCUE COMMITTEE
Save the Children
CDC
International Medical Corps
WATER
WEST AFRICAN TASK FORCE FOR THE CONTROL OF EMERGING & Re-EMERGING INFECTIOUS DISEASES
CEDEAO
ECOWAS
OOAS
WAHO
CDC
TDR For research on diseases of poverty
UNICEF • UNDP • World Bank • WHO
Public Health Institutes of the World
IANPHI

Only a selection of international responders is shown. There were many more.

Slide Credit: Laura Merson, IDDO

When the outbreak ended and organisations left the region, the data was scattered globally



product_id	time_id	customer_id	promotion_id	store_id	store_sales	store_cost	unit_sales
173	748	2094	54	1	4.29	1.8447	3
1119	748	2094	54	1	9.51	3.5187	3
1242	748	2094	54	1	7.92	2.8512	4
460	748	2094	54	1	6.44	2.7048	4
104	748	2094	54	1	11.67	3.9678	3
27	748	2094	54	1	7.95	3.8160	3
67	748	1277	54	1	7.44	2.9016	4
217	748	1277	54	1	2.72	0.8432	4
666	748	1745	54	1	2.67	1.3350	3
1122	748	1745	54	1	2.22	0.7770	3
899	748	1745	54	1	5.76	2.7072	3
423	748	1745	54	1	4.41	1.7640	3
1438	748	1745	54	1	2.74	1.0686	2

	ID	Age	Sex	Weight	Height	Married	Migrantstatus
1	1	26	1	132	60	0	Nonmigrant
2	2	65	0	122	65	0	Migrant
3	3	15	1	184	67	0	Nonmigrant
4	4	7	1	145	59	0	Nonmigrant
5	5	80	0	100	64	0	Migrant
6	6	43	1	NA	NA	0	Nonmigrant
7	7	28	1	128	67	1	Nonmigrant
8	8	66	1	154	60	1	Nonmigrant
9	9	45	0	166	NA	0	Migrant
10	10	12	0	164	60	1	Migrant

	Date of Observation	YEAR	QTR	Type of Observation	GC	GCD	GYD
1	1980:1	1980	1	ACTUAL	1702	218.7000	1894
2	1980:1	1980	1	ERROR	.	8.3333	0.2423
3	1980:1	1980	1	LOWER	.	185.7776	1843
4	1980:1	1980	1	PREDICT	.	210.3667	1893
5	1980:1	1980	1	STD	.	12.5457	25.9165
6	1980:1	1980	1	UPPER	.	234.9556	1944
7	1980:2	1980	2	ACTUAL	1705	198.2000	1901
8	1980:2	1980	2	ERROR	44.1085	-21.5633	-44.2237
9	1980:2	1980	2	LOWER	1715	195.1742	1895
10	1980:2	1980	2	PREDICT	1749	219.7633	1945
11	1980:2	1980	2	STD	17.5899	12.5457	25.9165
12	1980:2	1980	2	UPPER	1783	244.3524	1996
13	1980:3	1980	3	ACTUAL	1762	211.3000	1966
14	1980:3	1980	3	ERROR	9.8915	3.2521	13.1664
15	1980:3	1980	3	LOWER	1718	183.4588	1902
16	1980:3	1980	3	PREDICT	1752	208.0479	1953

NOTE: Table has been opened in browse mode. c:\Program Files\SAS

	1	2	3	4	5	6	7
Name	Sex	Fold	Pulse	Age	Clap	Exer	Smoke
Label							
Type	Factor	Factor	Number	Number	Factor	Factor	Factor
Format							
Levels	Female#...	L on R#,...			Left#,#...	Freq#,#...	Heavy#,...
1	Female	R on L	92	18.25	Left	Some	Never
2	Male	R on L	104	17.583	Left	None	Regul
3	Male	L on R	87	16.917	Neither	None	Occas
4	Male	R on L		20.333	Neither	None	Never
5	Male	Neither	35	23.667	Right	Some	Never
6	Female	L on R	64	21	Right	Some	Never
7	Male	L on R	83	18.833	Right	Freq	Never
8	Female	R on L	74	35.833	Right	Freq	Never
9	Male	R on L	72	19	Right	Some	Never
10	Male	R on L	90	22.333	Right	Some	Never

	1	2	3	4	5	6	7
Name	Sex	Fold	Pulse	Age	Clap	Exer	Smoke
Label							
Type	Factor	Factor	Number	Number	Factor	Factor	Factor
Format							
Levels	Female#...	L on R#,...			Left#,#...	Freq#,#...	Heavy#,...
1	Female	R on L	92	18.25	Left	Some	Never
2	Male	R on L	104	17.583	Left	None	Regul
3	Male	L on R	87	16.917	Neither	None	Occas
4	Male	R on L		20.333	Neither	None	Never
5	Male	Neither	35	23.667	Right	Some	Never
6	Female	L on R	64	21	Right	Some	Never
7	Male	L on R	83	18.833	Right	Freq	Never
8	Female	R on L	74	35.833	Right	Freq	Never
9	Male	R on L	72	19	Right	Some	Never
10	Male	R on L	90	22.333	Right	Some	Never

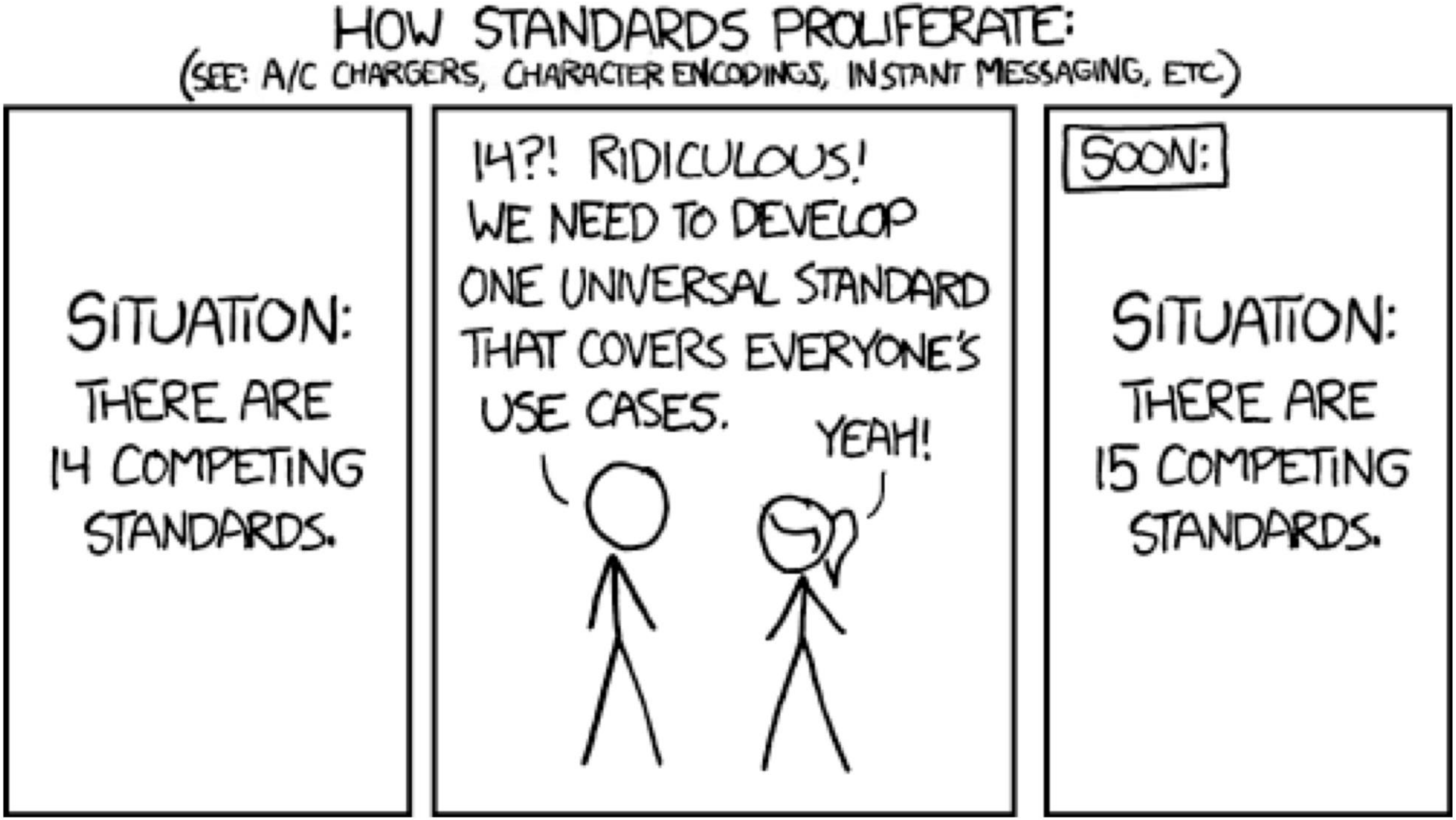
	ID	Age	Sex	Weight	Height	Married	Migrantstatus
1	1	26	1	132	60	0	Nonmigrant
2	2	65	0	122	65	0	Migrant
3	3	15	1	184	67	0	Nonmigrant
4	4	7	1	145	59	0	Nonmigrant
5	5	80	0	100	64	0	Migrant
6	6	43	1	NA	NA	0	Nonmigrant
7	7	28	1	128	67	1	Nonmigrant
8	8	66	1	154	60	1	Nonmigrant
9	9	45	0	166	NA	0	Migrant
10	10	12	0	164	60	1	Migrant

Only a selection of international responders is shown

Slide Credit: Laura Merson, IDDO

Initiative for Data Interoperability and Integration: Infectious Disease

- Data that characterise many of the factors influencing the progression of an outbreak are available, but remain isolated in siloes within the various domain- specific communities, often with their own domain-specific formats, vocabularies and ontologies.
- Availability of datasets from **industry, the research community, national public health surveillance, climate and environmental monitoring systems, health systems administration, social media feeds, and animal health services** will then be sought in order to understand how their integration can fill critical knowledge gaps across disciplines. Reports and lessons learned from previous infectious disease outbreaks have identified **clinical, genomic, demographic, pathogen and vector surveillance, communications, land-use, health administration, and environmental data** as powerful inputs to support planning and operationalising outbreak response. We can anticipate data in numerous formats such as tabular data in **spreadsheets, CSV, TSV, and/or plain text, geospatial point-wise data, geographic data, and a variety of XML and JSON dialects**. For the domains of interest, available ontologies will be sourced and compared to determine methods for integration and interchange.


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Source: xkcd.com

Metrics for Standards and Vocabularies

★ Someone’s text list (brain fart?) on the web

★★ Machine-readable lists

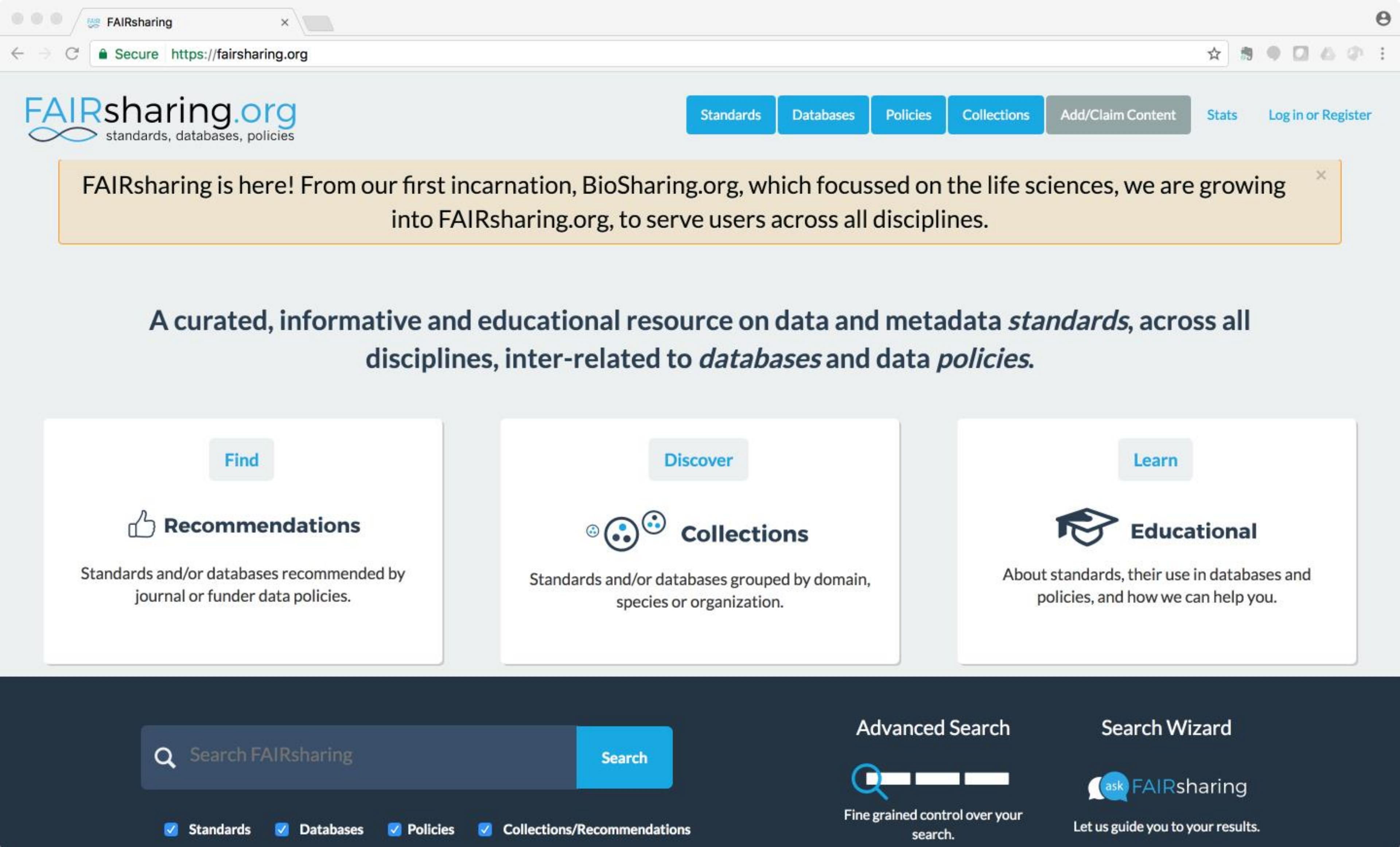
★★★ Non-proprietary lists of words with simple definitions

★★★★ Concept-based, community definitions, RDF, governed

★★★★★ Concept-based, RDF, linked, endorsed, multilingual

Slide credit: Lesley Wyborn

FAIRsharing



FAIRsharing is here! From our first incarnation, BioSharing.org, which focussed on the life sciences, we are growing into FAIRsharing.org, to serve users across all disciplines.

A curated, informative and educational resource on data and metadata *standards*, across all disciplines, inter-related to *databases* and data *policies*.

Find

Recommendations

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

Discover

Collections

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

Learn

Educational

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

Search FAIRsharing
Search
☑ Standards
☑ Databases
☑ Policies
☑ Collections/Recommendations
Advanced Search
Fine grained control over your search.
Search Wizard
ask FAIRsharing
Let us guide you to your results.

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.
- **Programme: https://conference.codata.org/2018_Goettingen_RDM/programme/**
- **Information/Registration:** http://www.eresearch.uni-goettingen.de/content/pre-rda-symposium

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

INTERNATIONAL DATA WEEK
IDW 2018

Gaborone, Botswana: 5-8 November 2018

ICSU CODATA

ICSU WORLD DATA SYSTEM

RDA RESEARCH DATA ALLIANCE

THUTO KE THEBE UNIVERSITY of BOTSWANA

ASSAf ACADEMY OF SCIENCE OF SOUTH AFRICA

INTERNATIONAL
COUNCIL
FOR SCIENCE

Thank you for your attention!

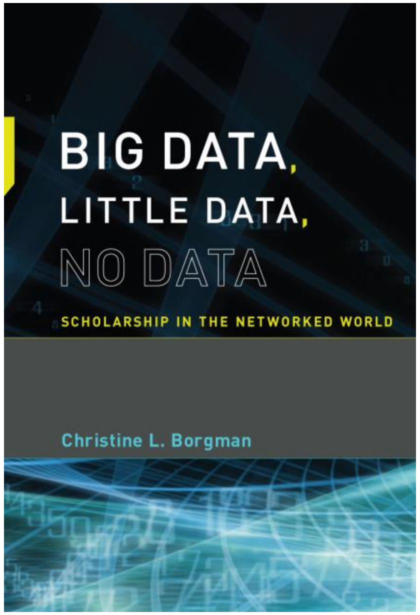
Simon Hodson
Executive Director CODATA
www.codata.org
http://lists.codata.org/mailman/listinfo/codata-international_lists.codata.org
Email: simon@codata.org
Twitter: @simonhodson99
Tel (Office): +33 1 45 25 04 96 | Tel (Cell): +33 6 86 30 42 59

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

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.**

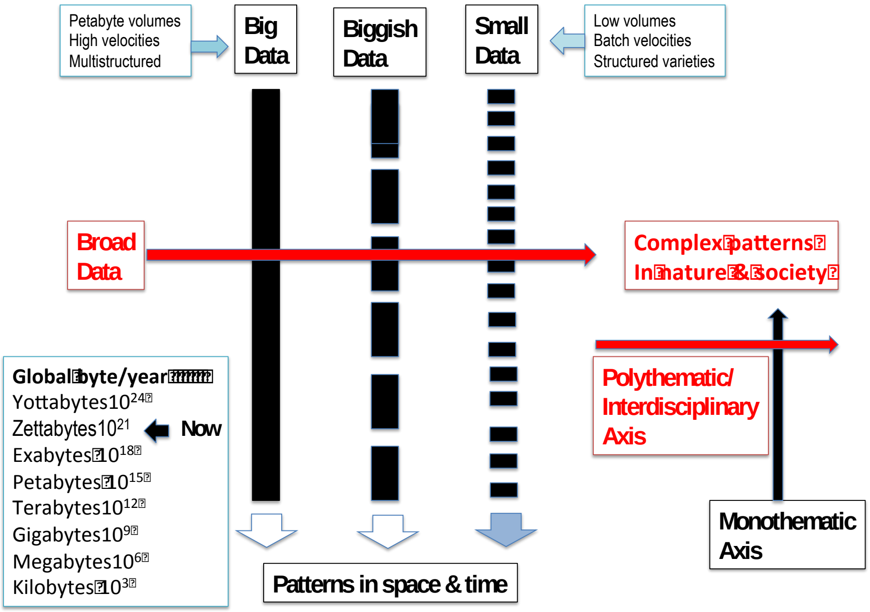
What are data? When are data?



BIG DATA, LITTLE DATA, NO DATA

SCHOLARSHIP IN THE NETWORKED WORLD

Christine L. Borgman



Petabyte volumes
High velocities
Multistructured

Big Data

Biggish Data

Small Data

Low volumes
Batch velocities
Structured varieties

Broad Data

Complex patterns In nature & society

Polythematic/ Interdisciplinary Axis

Monothematic Axis

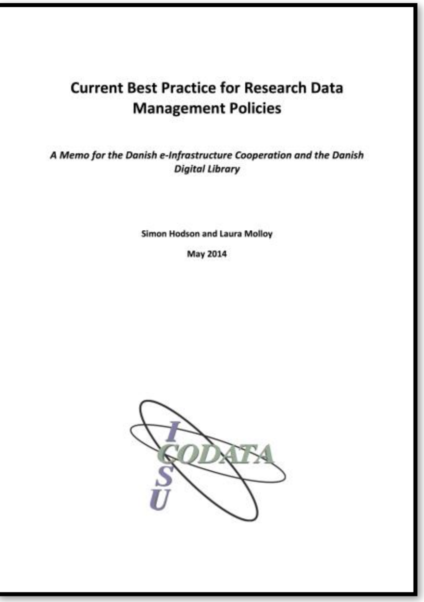
Patterns in space & time

Global byte/year

Yottabytes	10^{24}
Zettabytes	10^{21} ← Now
Exabytes	10^{18}
Petabytes	10^{15}
Terabytes	10^{12}
Gigabytes	10^{9}
Megabytes	10^{6}
Kilobytes	10^{3}

- Data is not the new oil.   They are far more valuable because they are renewable.
- **Big Data** poses many challenges and opportunities: how do we manage it and develop the data science to extract information?
- **Small data** poses many challenges, because often they require detailed human-crafted metadata.
- Often still our challenge is a **lack of data**, no data.
- Data can be very varied: poses challenges of storage, management, analysis.
- Many disciplines have raw data, processed data, science ready data or data products.
- Important to understand which of these can and / or should be preserved and made available.

Research data, public data…

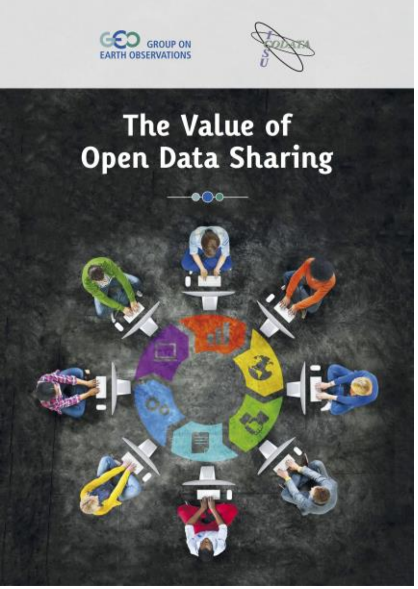


Current Best Practice for Research Data Management Policies

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

Simon Hodson and Laura Molloy

May 2014



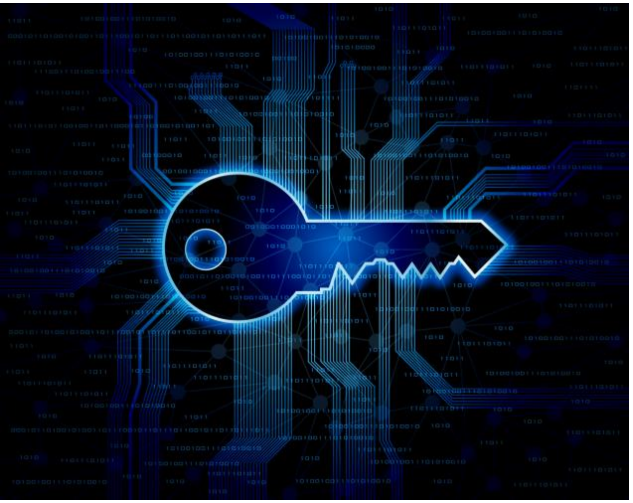
GROUP ON EARTH OBSERVATIONS

The Value of Open Data Sharing

- **Typology of data that are in scope of research data policies:**
  - Data underpinning research conclusions presented the literature must be published.
  - Significant data produced by research projects should be made available where possible.
  - Data resulting from major data creation projects for monitoring and research.
  - Data created by public activities where there are research and development opportunities.
  - Private sector data where there is a public good case or mutual benefit in data sharing.
- Research data should be as open as possible, as closed as possible.
- All research data should be FAIR.
- Framework for development of data policies; will be updated with additional resources in next 12 months by CODATA Data Policy Committee.

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.**

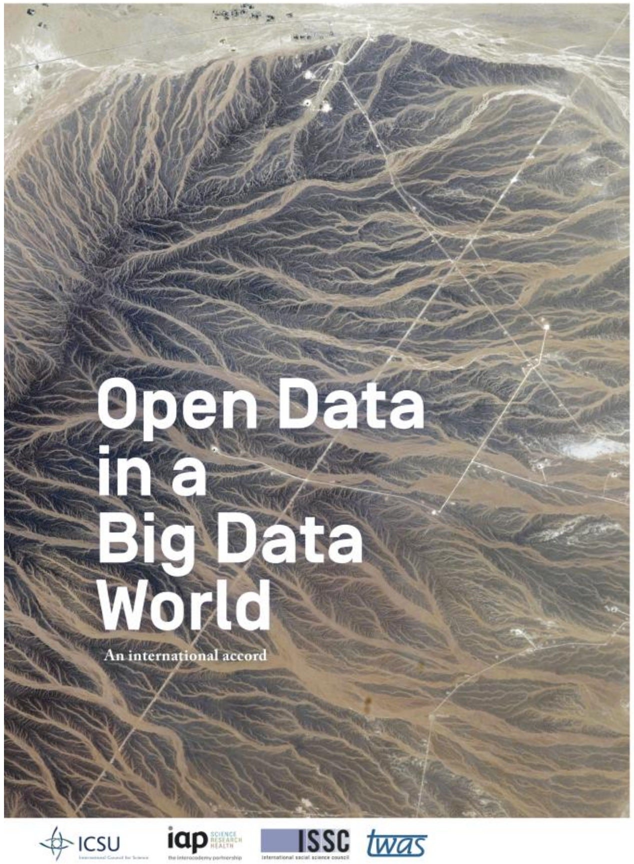


Criteria for appraisal, selection, preservation?

- Need for more work on criteria and guidelines for appraisal, selection, preservation.
- A lot of the criteria has to come from the disciplines, but informed by some general principles.
  - Major data collection/creation exercises with **evident multiple uses** (census, Hubble etc).
  - Unrepeatable observations, measurements in nature or society? **(preserve and publish)**
  - Data created by in vitro experiments that can be reproduced and for which the instruments are being improved **(perhaps very limited reasons for preserving)**
  - Data collected for a given public or private purpose (traffic management, customer relations, ships logs) but which could be used for research…
- **Need for collaborative approach (with researchers) to clarifying the criteria to keeping, publishing and discarding data.**
- **While being mindful of potential for other extra- and inter-disciplinary uses.**
- Qualified by the knowledge that some disciplines have to discard data because of the sheer volume.
- What is important is that we start actively exercising these processes.
- DCC Guidelines: http://www.dcc.ac.uk/resources/how-guides/appraise-select-data
- NERC Data Value Checklist: http://www.nerc.ac.uk/research/sites/data/policy/data-value-checklist/

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**



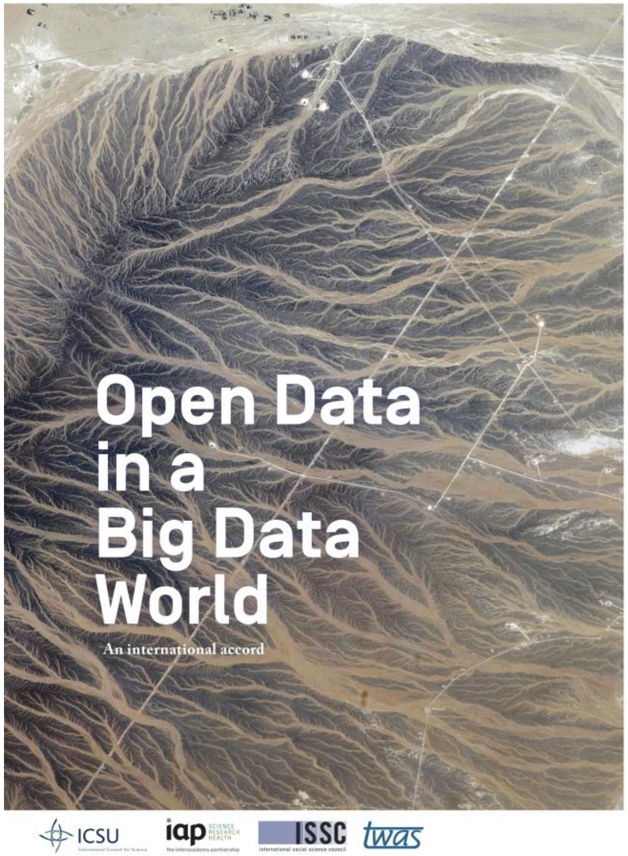
Open Data in a Big Data World

An international accord

ICSU iap ISSC twas

Framework for Regional, National and Institutional Data Strategies

- **National / Institutional Open Science and FAIR Data Strategy**
  - Consultative forum, stakeholder engagement.
- Open data **policies and guidance** at national and institutional level.
  - Clarify the **boundaries of open** (particularly privacy, IPR).
  - Clarify the data in **scope**, guidelines on selection.
- Develop **incentives and reward** systems.
  - 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**:
  - Essential **data skills for researchers**.
  - Develop skills and competencies for **data stewards, data scientists**.



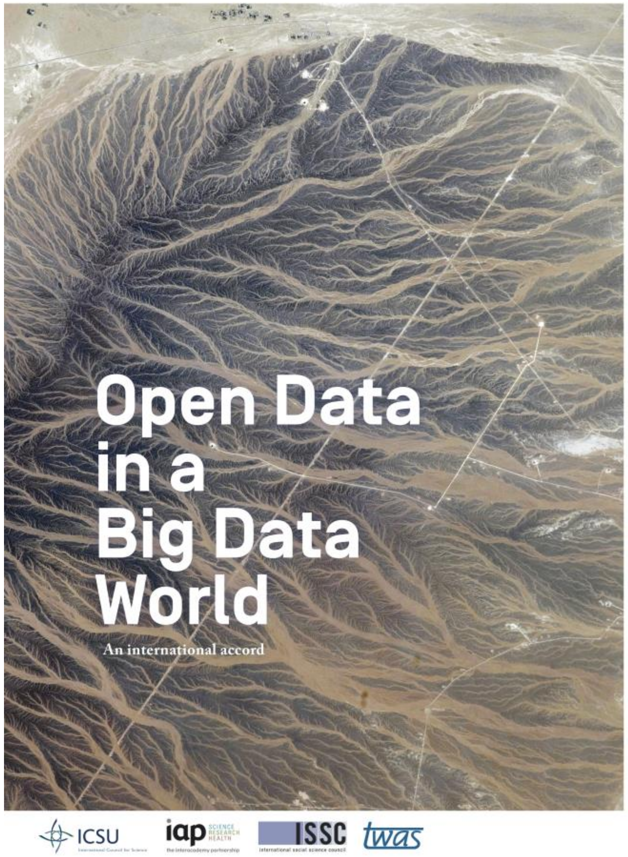
Open Data in a Big Data World

An international accord

ICSU iap ISSC twas

Framework for Regional, National and Institutional Data Strategies

- Scope, roadmap and implement data **infrastructure**.
  - Key **components of national and regional infrastructure** (network / NREN, economies of scale for storage and compute).
  - Development **of regional, national and institutional infrastructure(s)** for research collaboration and data stewardship/RDM, generic research platforms/environments, trusted digital repositories.
  - **Collaborative infrastructures** for certain research disciplines, nationally, regionally to pool expertise and lower costs.
  - **International infrastructure / data ecosystem components**: permanent identifiers, metadata standards.



Open Data in a Big Data World

An international accord

ICSU iap ISSC twas

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

CODATA RDA SCHOOL

OF RESEARCH DATA SCIENCE

TRIESTE 2017

The Abdus Salam
International Centre
for Theoretical Physics

twas