Big Data, Data and CODATA

Simon Hodson

Executive Director CODATA

www.codata.org

execdir@codata.org

@simonhodson99

Big Data and its Challenges

- What do we mean by Big Data?

- Three Vs (Volume, Velocity and Variety) are a good definition (any more is simply marketing).

- Data is 'Big' when it forces us into new ways of dealing with it.

- These pose significant technical, economic and cultural challenges.

- Many examples of major projects producing huge volumes of data, at great rates… Sometimes also with great complexity.

- **For science, Big Data is about Big, but its also about data… (Chaitan Baru).**

  - Disciplinary and interdisciplinary challenges.

  - Knowledge and validity (how do we demonstrate proofs effectively).

  - Education and training.

 - Ethics (particularly important with data relating to human subjects).

ALMA
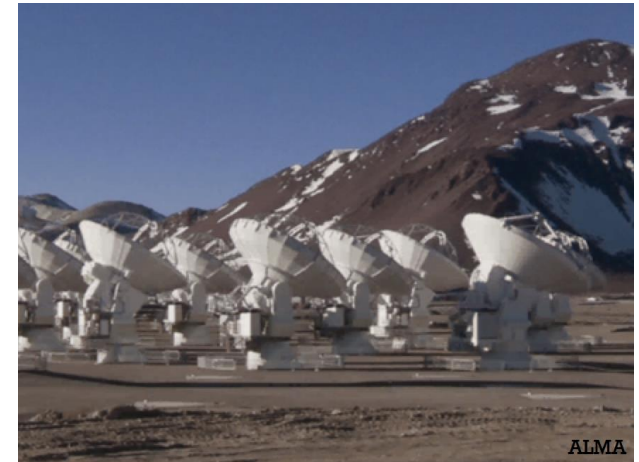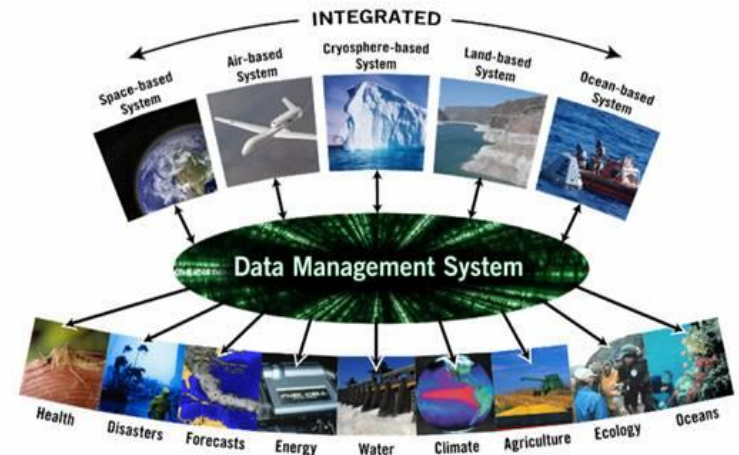
The Data Revolution…

- Creating a world that counts… http://www.undatarevolution.org/

- Independent Expert Advisory Group on the Data Revolution for Sustainable Development (IEAG)

- Report issued in early November.

- Huge growth in data: not just volume but new types of data, new sources of information.

- Huge opportunities for understanding and monitoring sustainable development.

- Call to arms to use the data at our disposal!

- Challenges and risks to be addressed:

  - Challenges of data inequality

  - **Challenges of data integration (traditional vs new)**

  - Risks relating to privacy, trust, data quality.

A WORLD THAT COUNTS

MOBILISING THE DATA REVOLUTION FOR SUSTAINABLE DEVELOPMENT

DATA REVOLUTION GROUP
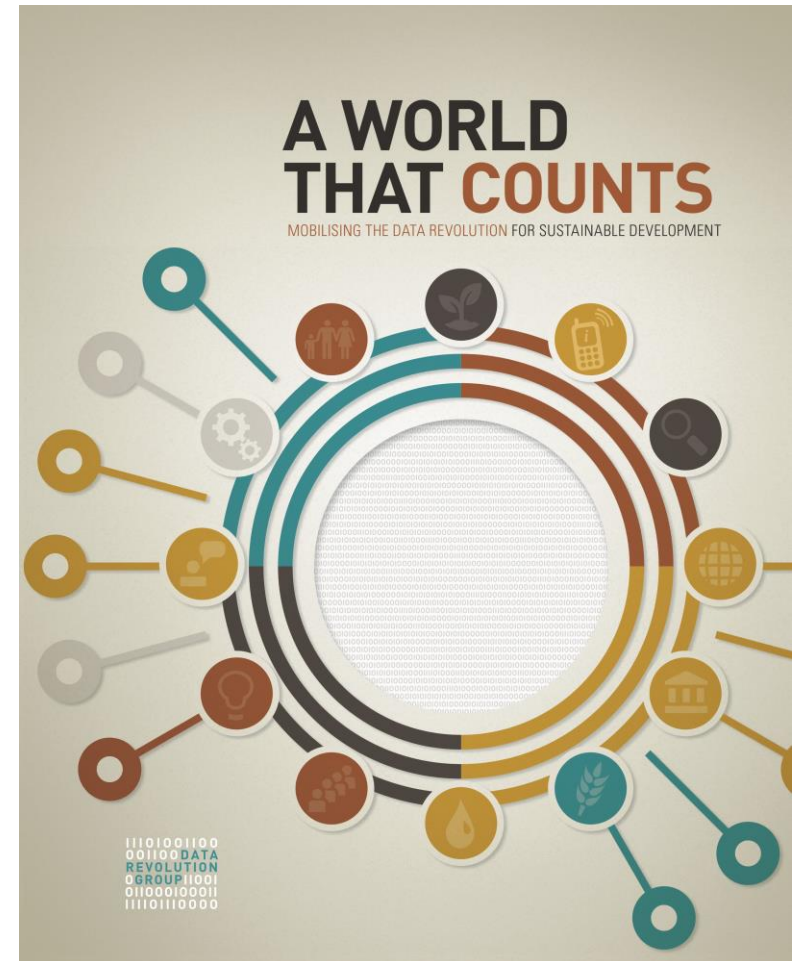
photos: www.dawide.com

futurearth
research for global sustainability

To provide the knowledge required for societies in the world to face risks posed by global environmental change and to seize opportunities in a transition to global sustainability

Ten-year interdisciplinary research programme;
Sponsored by ICSU, ISSC, Belmont Forum, UNESCO, UNEP

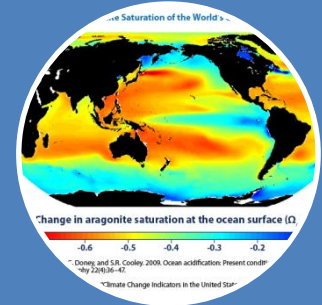# Future Earth Data Integration

Existing Data:

Existing Knowledge

New Observations

Models

Integrated and interdisciplinary datasets, indicators, visualizations, scenarios, information…

Social, economics, ecosystems, geophysics

Availability (access, regulations, ethics); heterogeneity (standards, units); interoperability; quality.

CODATA: Committee on Data for Science and Technology
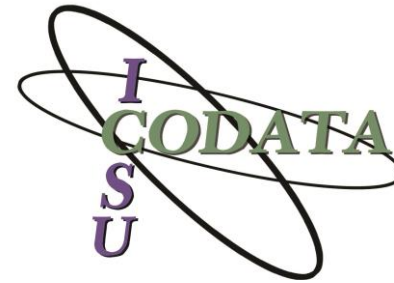
ICSU's Mission ⟶ **CODATA's Mission**

*"Strengthen international science for the benefit of society **by promoting improved scientific and technical data management and use**."*

CODATA History and Mission

By the early **1960s** a number of scientific leaders began to realize that this deluge of data was swamping the traditional publication and retrieval mechanisms, and that there was a danger that much of it would be lost to future generations. When several of these leaders got together and agreed that an organized international effort was needed to improve the management and preservation of scientific data and to facilitate coordination among interested groups throughout the world, **the creation of CODATA was the outcome.**

CODATA @ 45 Years: the story of the ICSU Committee on Data for Science and Technology

Open Data Policies

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

Expert Report
on data policies
for Danish e-
Infrastructure
Group

Workshop on
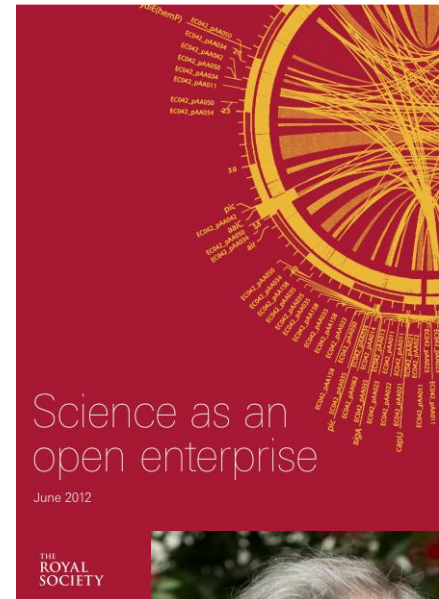data policies for
ICSU Europe
planned for
Spring 2015

COUNCIL OF FINNISH ACADEMIES

GEO GROUP ON
EARTH OBSERVATIONS

GEO-X Plenary & Geneva Ministerial Summit
Side Event

Implementation of
Data Sharing Principles

Monday, 13 January 2014, 14:30-18:00
Salle 5+6, Level 3, CICG

Substantial input to
ICSU Report on
Statement on Open
Access and Metrics
http://bit.ly/icsu-OA-
statement

ICSU
International Council for Science

Science as an
open enterprise

June 2012

THE
ROYAL
SOCIETY

Leading role in GEO DSWG
and DMP TF
http://bit.ly/GEO_DSPs

New CODATA President
Geoffrey Boulton, FRS
Chair of *Science as an
Open
Enterprise* Report

GEO

CODATA-ICSTI Task Group
Data Citation, Standards and Practices

For Attribution

Workshop and Report:
http://bit.ly/for_attribution

Out of Cite, Out of Mind

http://bit.ly/out_of_cite

DATA CITATION STANDARDS AND PRACTICES
(http://www.codata.org/taskgroups/TGdatacitation/index.html)

Jan Brase, Co-Chair   Sarah Callaghan, Co-Chair   Christine Borgman, Co-Chair

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

DC^1

Background and Developments:
http://bit.ly/data_citation_principles
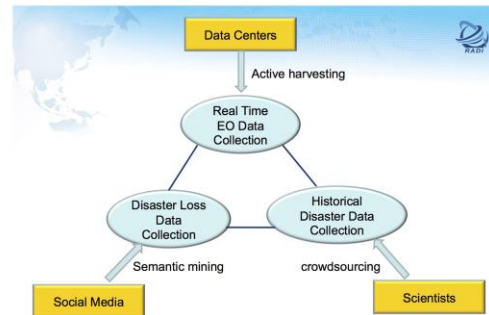
Data Citation Principles

Challenges in Data Science

Advancing Informatics for Microbiology

Collaboration around interoperability and knowledge extraction issues; research and training workshops: http://www.codata.org/task-groups/advancing-informatics-for-microbiology
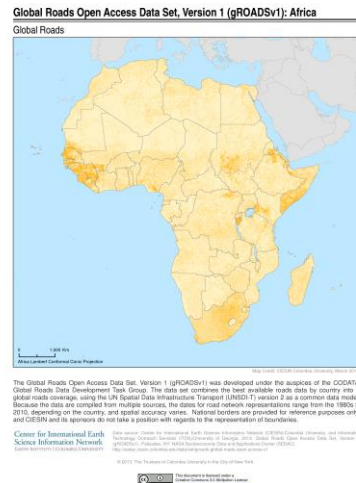
LOD Global Disaster Data

Preparing White Paper on Use of LOD for Disaster Data: http://www.codata.org/task-groups/linked-open-data-for-global-disaster-risk-research

Data Science Journal

New platform and partnership for CODATA's OA Data Science Journal soon to be announced!

Global Roads Data

Review of Global Roads Data Development Methodologies: http://bit.ly/globalroads-methods

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

CODATA-ICSU Workshop: http://www.codata.info/Nanomaterials/Index-agenda-Nanomaterial.html
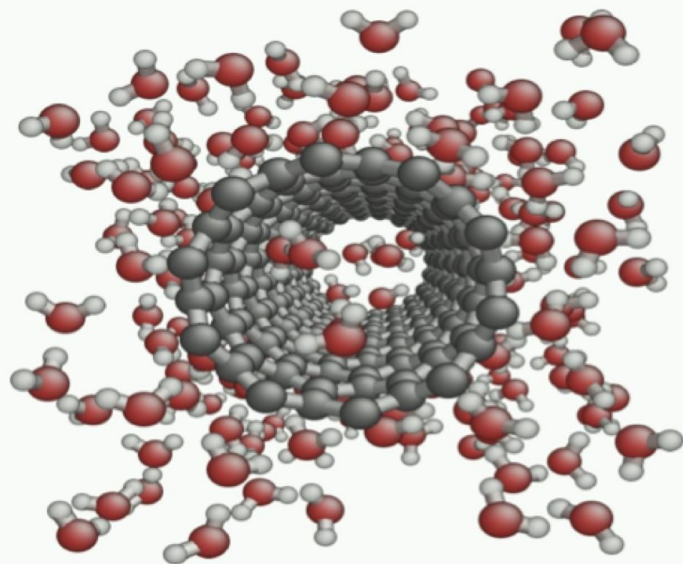
CODATA WG on the Description of Nanomaterials: http://www.codata.org/nanomaterials

Draft Uniform Description System: http://www.codata.org/uploads/Nanomaterials_Framework_for_Uniform_Description_System_2013-10-03.pdf

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

International Collaboration

SciDataCon 2014 — International Conference on Data Sharing and Integration for Global Sustainability

SciDataCon 2014, 2-5 Nov, New Delhi: co-organised with ICSU-WDS http://www.scidatacon2014.org/

Motivated by the conviction that the most significant research challenges – and in particular the pressing issues relating to global sustainability – cannot be properly addressed without due attention to various issues relating to data.

SciDataCon aims to make explicit the connection between concrete and specific challenges to ensure environmental protection, economic prosperity and social well being along with the research questions that arise from these challenges and the necessary role of data policy, management and analysis in addressing these research questions.

CODATA Big Data Workshop
Beijing, 8-9 June 2014

- Aim and Objectives

    - to create an international and cross-disciplinary **awareness** of the wide and diverse activities underway in the science-related Big Data sphere;

    - to establish and foster ongoing **coordination and collaboration** for international, interdisciplinary research programmes, such as IRDR and Future Earth, on Big Data issues;

    - to identify ways in which a **CODATA Working Group on Big Data for International Science** programmes can contribute.

- Outcomes

    - **Statement of Recommendations on Big Data for International Sciences Programmes.**

    - White Paper and Editorial

    - CODATA Working Group on Big Data

Big Data for International Scientific Programmes: A Statement of Recommendations and Actions

3 Replies

There is little doubt that Big Data is a hot topic. Yet while the significance for Big Data may be demonstrable in certain research areas, there is also a lot of hype (particularly in relation to commercial applications), and its corollary, therefore, scepticism. Convened by CODATA and co-sponsored by a number of important international organisations, the *International Workshop on Big Data for International Scientific Programmes* took place on 8-9 June in Beijing, and aimed to shed more considered light on the potential role of Big Data in such international and interdisciplinary research activities.

Professor GUO Huadong, President of CODATA gives the opening keynote

http://codata.org/blog/2014/06/16/big-data-for-international-scientific-programmes-a-statement-of-recommendations-and-actions/

Big Data: statement of recommendations and actions

**Big Data for International Scientific Programmes:
Challenges and Opportunities
A Statement of Recommendations and Actions
Beijing 9 June 2014**

Recommendatons

- Respond to the importance of Big Data for international scientific programmes

- Exploit the benefits of Big Data for society

- Improve understanding of Big Data through international collaboration

- Promote universal access to Big Data through global research infrastructures

- Explore and address the challenges of Big Data stewardship

- Encourage capacity building and skills development in Big Data science

- Foster development of policies to maximise exploitation of Big Data

Actions for a CODATA Working Group

- Produce case studies in Big Data for international scientific programmes

- Promote sharing of Big Data solutions across scientific disciplines

- Research policy, ethical and legal issues for Big Data

- Research into stewardship and sustainability challenges for Big Data

http://bit.ly/Big_Data_Recommendations

Capacity Building

CODATA Training in Big Data Science
Beijing, 4-20 June 2014
http://bit.ly/CODATA-China_Training_2104-Call

Training Workshop on Open Data, Kenya, Jomo Kenyatta University of Science and Technology
http://bit.ly/codata-training-jkuat

International Workshop on Open Data for Science and Sustainability in Developing Countries

- Convened by CODATA Task Group, *Preservation of and Access to Scientific and Technical Data **in/for/with** Developing Countries*.

- Partners included UNESCO; Kenyan Ministry of Information, Communications and Technology; Jomo Kenyatta University of Agriculture and Technology; GEO, ICSU-WDS, RDA.

- Two day training workshop covered:

  - Data policies; legal and licensing issues.

  - Data management principles and planning

  - Data sources

  - Various approaches to data integration and analysis, particularly geospatial and EO data.

- Roughly 70 participants in both parts of the workshop.

International Workshop on Open Data for Science and Sustainability in Developing Countries

- Strong endorsement for the workshop from Kenyan Cabinet Secretary and from local universities and research institutes.

- Cabinet Secretary Dr. Fred Matiang'i: announced data centre to be established at JKUAT

- Called on CODATA and other international organisations to 'become more visible in education and capacity-building, by developing science and educational programs and activities that focus on data and information' in developing countries.

- Discussed Open Data Principles, the application of such principles in developing countries and their relationship to scientific objectives supporting SDGs.

- Developed a set of ***Principles of Preservation of and Open Access to Research Data in Developing Counties***

- Prepared outline of ***Guidelines*** for their implementation.

- Particular concerns for credit and periods of privileged access; concerns about giving away intellectual assets for free…

CODATA Training in Big Data Science
Beijing, 4-20 June 2014

- Training Workshop for young researchers, research leaders and managers of research institutes: http://bit.ly/CODATA-China_Training_2104-Call

- Broad Data Science programme covering policy, theory, data management, analysis.

- 21 participants from 10 countries (including Kenya, Uganda, Tanzania, South Africa)

- Built on previous workshop in 2012.

- **Ambition: to make this a regular series!**

- **Develop a regular CODATA Training Series in various regions, in partnership with local institutions.**

RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World

- Initiative to collaborate with RDA and augment existing CODATA activities: appeal for input from PASTD, ECDP and others.

- Proposal currently under consideration within RDA: http://bit.ly/CODATA-RDA_Data_Science_Summer_Schools

- Establish framework to run a series of Summer Schools in Data Science and Cloud Computing in the Developing World.

  - Arrange funding for an initial period for the school to run (five years).

  - Organise partnerships with Developing World institutions.

  - Determine the best curriculum for the school in collaboration with others.

  - Arrange how the materials can also be delivered online.

- **Currently exploring funding opportunities and looking for partners and sponsors.**

Curriculum Framework for (Big) Data Science?

- **Big Data MOOCs are emerging e.g. https://bigdatacourse.appspot.com/preview**

- **Currucframework in SCCID report**

- **Programme for CODATA training workshop**

- What are the key curriculum elements?

    - Introduction - What is Big Data? What are its features?

    - Big Data Storage – formats – scaling

    - Tools and techniques for handling Big Data [4-8 lectures]

    - Standards for Big Data

    - Semantics for Big Data organization and retrieval

    - Data models and services

    - Domain based data

    - Open Government Data

    - Big Data Licensing: access, use, reuse modalities

    - Data and text mining

    - Approaches to structured and unstructured data

    - Case studies, Big data projects and alliances

    - And more…

International Big Data Training Workshop at Indian Statistical Institute, Bangalore, March 2015

- Led by experts at ISI and drawing on CODATA international community.
- Using the model developed for the Beijing and Nairobi workshops, but targeting students from India and SAARC countries.
- Bringing experts to the students.
- **Looking for sponsors of travel for international experts.**
- **Looking for experts interested in participating.**
- **Data publishing? Big data and scientific proof? Big data and scholarly communication? Big data and research ethics?**
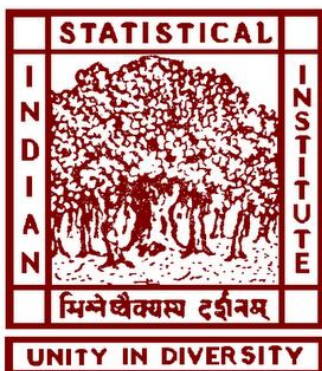
Thank you for your attention!

Simon Hodson
Executive Director CODATA
www.codata.org/blog
Email: execdir@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