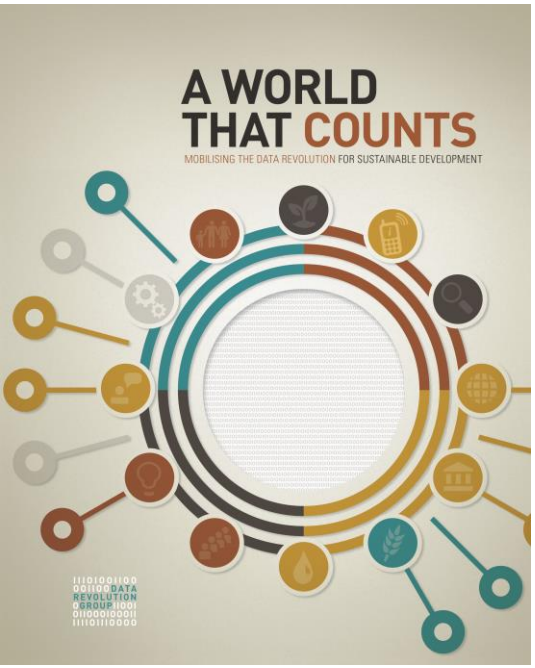
Institute for Data Science Seminar Series
Carleton University, Ottawa, Canada
29 November 2016

Open Data and the Data Revolution: Challenges and Opportunities for Global Research

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

Data Revolution: A World that Counts!



- **Creating a world that counts:** Mobilising the Data Revolution for Sustainable Development.
- To meet the new sustainablity goals '*there is an urgent need* ***to mobilise the data revolution for all people and the whole planet in order to monitor progress****, hold governments accountable and foster sustainable development.'*
- *Without immediate action****, gaps between developed and developing countries, between information-rich and information-poor people, and between the private and public sectors will widen****, and risks of harm and abuses of human rights will grow.*
  - Data quality and integrity
  - Data disaggregation (no-one should be invisible)
  - Data timeliness
  - Data transparency and openness
  - Data usability and curation
  - Data protection and privacy
  - Data governance and independence
  - Data resources and capacity
  - Data rights

THIS IS THE REVOLUTION

Indonesia is one of the most social-media dense countries in the world today. Indonesians tweet about a range of topics, including the cost of living. A project by UN Global Pulse, the Indonesian Ministry of National Development Planning and the World Food Programme found public tweets mentioning food prices closely approximate official figures, leading to the development of a technology that extracts daily food prices from public tweets to generate a near real-time food price index. This data mining approach could be adapted to other food items and locations, not just leveraging Twitter but other crowd-sourced and social data sources.

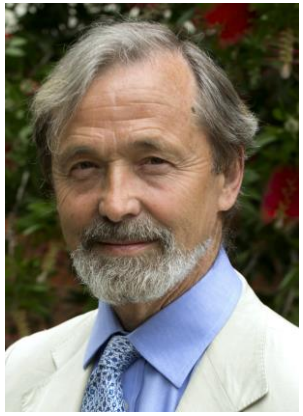
Source: UN Global Pulse (http://www.unglobalpulse.org/nowcasting-food-prices)

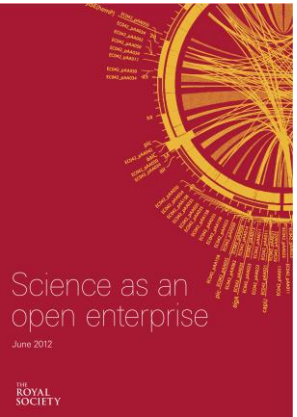
Data Revolution: how can we improve … with open data?

- GODAN-ODI Report: improving agriculture, food and nutrition with open data.
- *'Although the amount of data openly available is constantly increasing,* ***there are still challenges related to data management, licensing, interoperability and exploitation. There is a need to evolve policies, practices and ethics around closed, shared, and open data.****'*
- **Enabling more efficient and effective decision making** > lowers cost of accessing information and underpins tools that farmers themselves can use.
- **Fostering innovation to benefit everyone** > an opportunity that must not be missed for creating new businesses and jobs in 'new data-powered innovation ecosystems'.
- **Driving organisational and sector change through transparency** > open data is essential to understanding complex systems, interventions, targets, change.
- **Availability is not enough** > essential that the data be interoperable and machine-readable.
- Problem oriented and solution-based data strategies.
- Develop infrastructure and human capacity.

CODATA: Committee on Data of the International Council for Science

CODATA President Geoffrey Boulton, FRS





Royal Society Report: Science as an Open Enterprise

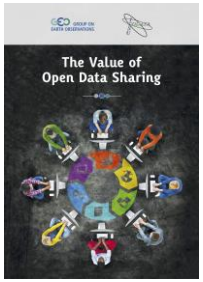
- CODATA President, Geoffrey Boulton, was lead author and chair of **Royal Society Report: Science as an Open Enterprise.**
- Identifies challenges and opportunities for science systems, technical and human.
- Fundamental methodological issues for reproducibility and transparency.
- **Publications and data should be Intelligently Open and available concurrently.**
- **Report with very significant impact: G8, H2020**
- **CODATA**Established by the International Council of Science to address issues of data availability and quality.
- Remit has broadened over the years.
- New Executive Committee: includes members from Kenya and South Africa, will co-opt a member from Latin America.
- **Increased orientation towards playing a coordinating role on national and regional Open Science strategies.**

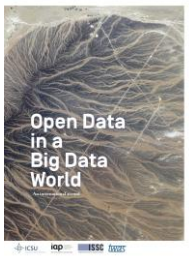
CODATA

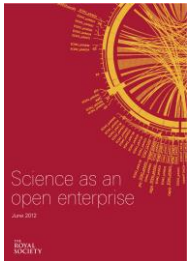
Principles, Policies and Practice

DC1 Data Citation Principles

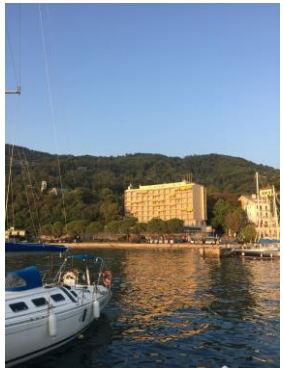
ICSU International Council for Science







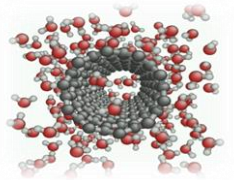
Capacity Building

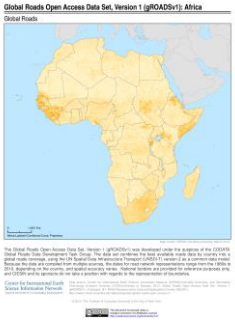


UNESCO

The Abdus Salam International Centre for Theoretical Physics

Frontiers of Data Science





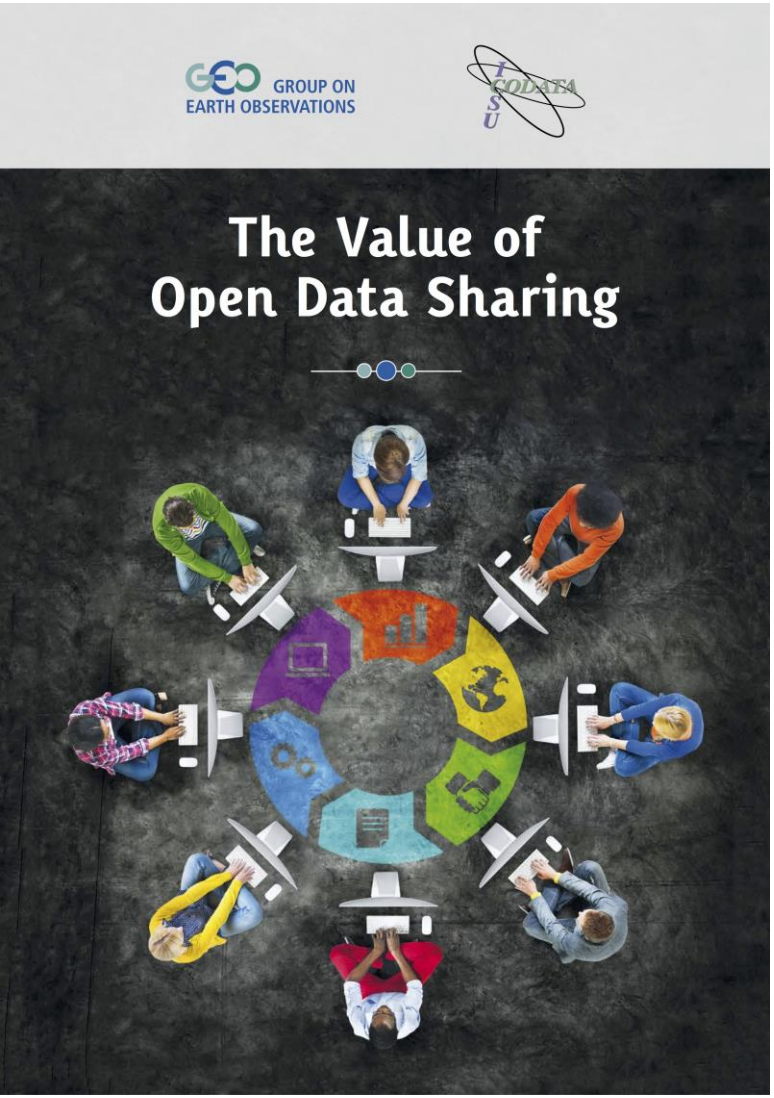
ubiquity press open scholarship

Data Science Journal

IDW 2016, 11-17 Sept, Denver, CO.

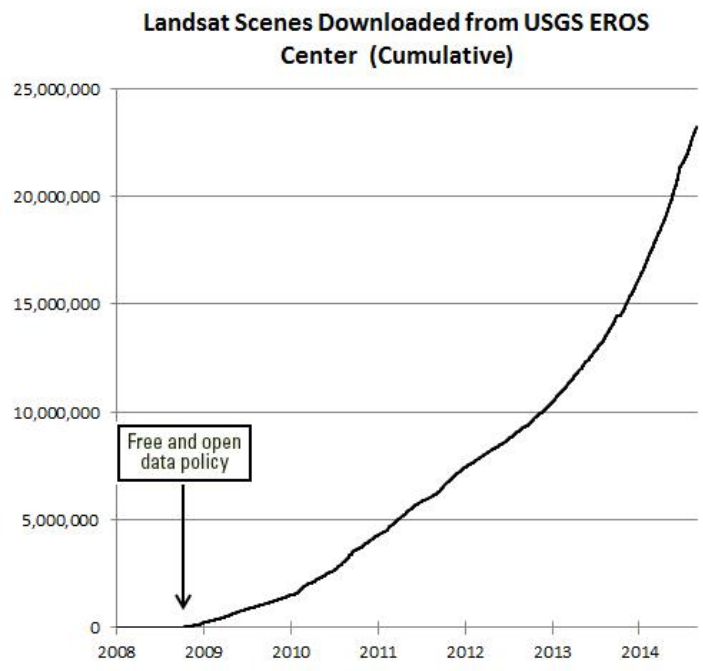
INTERNATIONAL DATA WEEK 2016
WWW.INTERNATIONALDATAWEEK.ORG

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.**

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.**

Data Policies

OECD Principles and Guidelines for Access to Research Data from Public Funding

http://bit.ly/oecd_principles

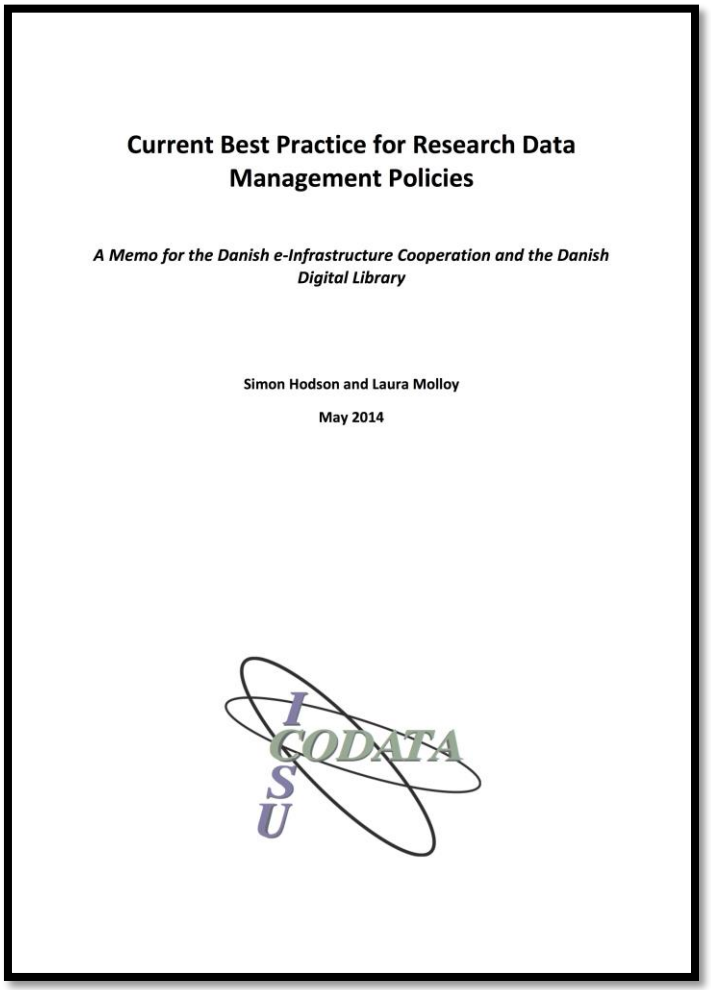
Access to research data increases the returns from public investment in this area; reinforces open scientific inquiry; encourages diversity of studies and opinion; promotes new areas of work and enables the exploration of topics not envisioned by the initial investigators.

Science and Technology Ministers called on the OECD in 2004 to develop a set of guidelines based on commonly agreed principles to facilitate cost-effective access to digital research data from public funding.

Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based.

Resources: Current Best Practice for Research Data Management Policies

- **Expert report commissioned by CODATA member.**
- Provides comprehensive summary of best practice in funder data policies.
- Identifies key elements to be addressed:
  1. Summary of policy drivers
  2. Intelligent openness
  3. **Limits of openness**
  4. **Definition of research data**
  5. **Define data in scope**
  6. **Criteria for selection**
  7. Summary of responsibilities
  8. Infrastructure and costs
  9. DMP requirements
  10. Enabling discovery and reuse
  11. Recognition and reward
  12. Reporting requirements, compliance monitoring
- Zenodo: http://dx.doi.org/10.5281/zenodo.27872



Implementation Guidelines for the Legal Interoperability of Research Data

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

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

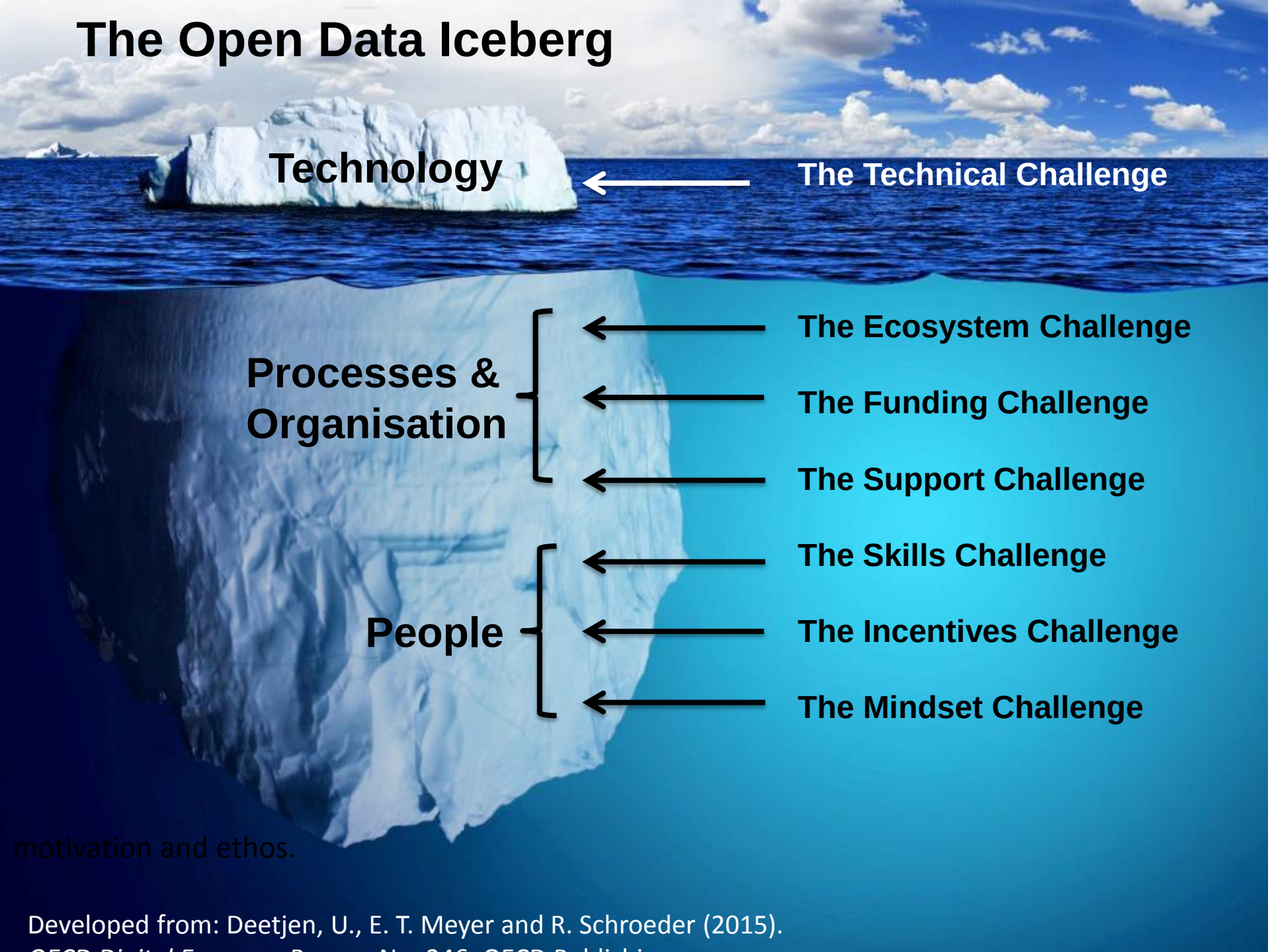
The Case for Open Data in a Big Data World

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

African Open Data Platform Initiative ICSU-CODATA

- Proposals for Open Data Platform initiatives, Africa and Latin America and Caribbean.
- Holistic 'science systems' approach: policies, procedures, incentives, data infrastructure, scholarly communications, skills and training.
- **Keystone is to establish an Open Data Platform with a coordinating role.**
- Pilot initiative funded by Department of Science and Technology in South Africa: nearly 500K euros over three years.
- Implemented by staff from South African Academy of Sciences, under direction from ICSU-CODATA.
- **Currently undertaking preparatory study chart the data landscape and build partnerships**

The Open Data Iceberg



motivation and ethos.

Developed from: Deetjen, U., E. T. Meyer and R. Schroeder (2015).

Building the Initiative

Establish African Open Data Forum / Platform

Co-design African Open Data Policies

Develop Incentives Frameworks

Develop Research Data Science Training

African Research Data Infrastructure Roadmap

Activities require low funding for coordination, secondment, contributions in kind and evaluation.

Activities require higher investment for coordination, co-design implemenatation and evaluation.

Funded Research Data Infrastructure Initiatives

Funded, co-designed transdisciplinary research projects

Research Data Infrastructure Roadmaps

- **Research priorities and gap analysis.**
- **Ecosystem:** what is the provision of RDIs for particular disciplines through national and international initiatives?
- **Role of Research Institutions:** Is lifecycle support and long tail being supported in institutions.
- RDIs are not just hardware, but 'part of a research ecosystem', so must address: governance; training, personnel and career structures' sustainable funding; access and outreach to national, public and commercial partners.
- **Roadmap for Data Infrastructure**
  - Co-design to meet national needs and priorities.
  - Research priorities, opportunities for shared infrastructures, examples of good governance and sustainable funding models.
  - **Sustainable Business Models for RDIs**

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)

The Challenge: Sustainable Business Models for Data Repositories

- Research funder policies – quite rightly – mandate data stewardship.
  - OECD Principles and Guidelines, 2007
  - G8 Science Ministers Statement, 2013
  - Major funders in US, UK, EC Horizon 2020 data policy etc.
- 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.
- Witnessing Innovation
  - Changes in funding / business models (ADS, TAIR; DANS, ICPSR)
  - Innovative business models (Dryad, FigShare)

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

- Questions to address:
  1. How are data repositories currently funded?
  2. What innovative income streams are available?
  3. What means of restraining costs are available?
  4. How do income streams match willingness/ability to pay of various stakeholders?
  5. How do income streams/willingness to pay fit together into a **sustainable** business model?
- Builds on previous work of RDA-WDS Interest Group: http://dx.doi.org/10.5281/zenodo.46693
- Broader landscape survey of current funding models, May-Sept 2016.
- Focus group on innovative income streams and on cost restraint, workshop Nov 2016.
- Micro and macro economic analysis of business models, Nov 2016-Mar 2017.
- Test business models with stakeholder groups, workshop April 2017.
- Policy recommendations based on concrete business model options, April-June 2017.

OECD BETTER POLICIES FOR BETTER LIVES

ICSU WORLD DATA SYSTEM

RDA RESEARCH DATA ALLIANCE

An Open Research Data Strategy for Poland

- Collaboration on a national workshop to develop a national open research data / open science strategy for Poland.
- CODATA leads met earlier this year with representatives from Ministry of Science and Education and with Open Science Centre to plan a workshop for Feb / March 2017.
- Draws strongly on the approach of the accord.
- **Stakeholders and Responsibilities:** governments/funders, universities and research institutions, institutional libraries, national academies and learned societies, national and international research and data infrastructures, publishers and journal editorial boards.
- **Working Groups on Enabling Practices:** boundaries of open, normative values (sharing, timeliness), non-restrictive reuse and TDM, incentives, interoperability, sustainability of data infrastructure, data literacy.

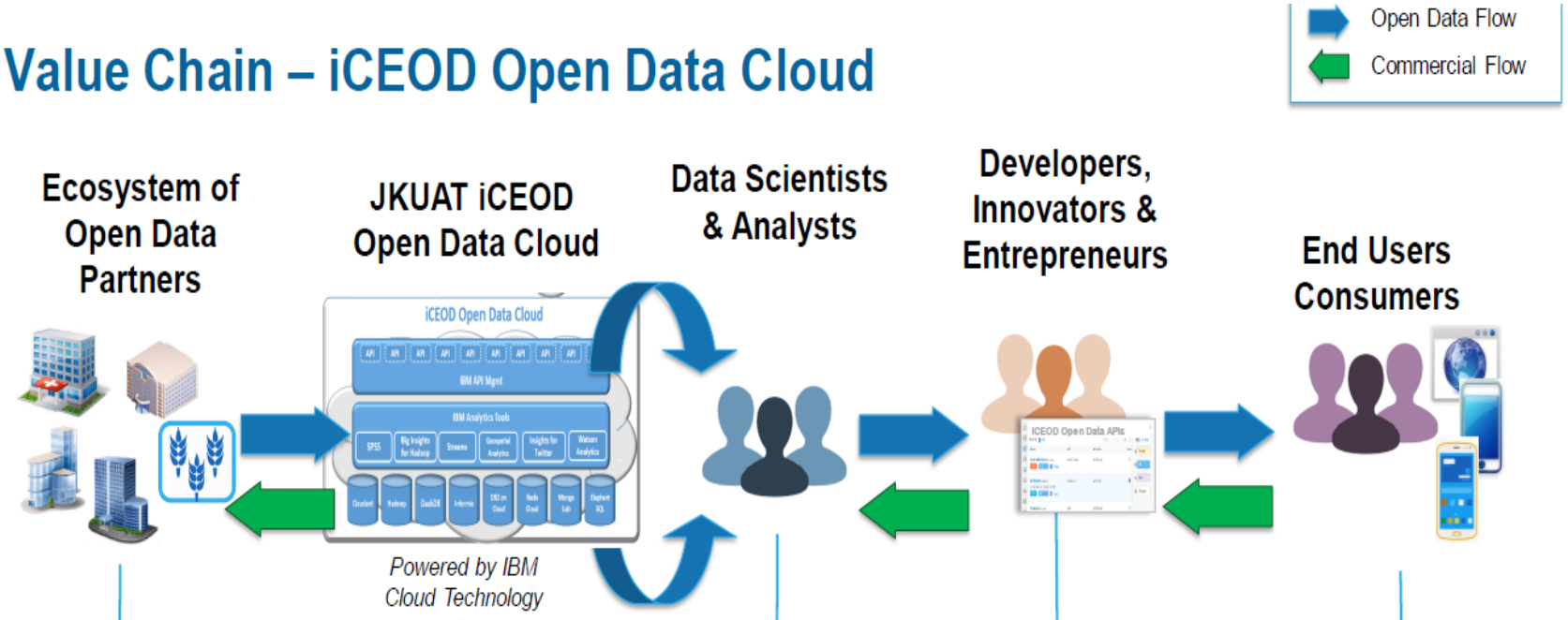
OPEN SCIENCE PLATFORM

CODATA in Kenya

- International workshop on open data for science in developing countries, UNESCO, Nairobi, August 2014.
- Strong endorsement for the workshop from Kenyan Cabinet Secretary and from local universities and research institutes.
- Cabinet Secretary Dr. Fred Matiang'i: 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.
- Announced data centre to be established at Jomo Kenyatta University of Agriculture and Technology.
- **'JKUAT has now established an ICT Centre of Excellence and Open Data (iCEOD) that was part of the Nairobi-CODATA conference recommendation'**
- Working with CODATA on data management policies and development of iCEOD and data system: http://www.codata.org/membership/national-members/kenya

iCEOD Value Chain: Data and Society Agriculture and Nutrition

Value Chain – iCEOD Open Data Cloud



Open Data from iCEOD and their partner ecosystem including
- Hospitals
- Research Partners
- Agriculture Partners
- Governments

JKUAT Staff and PhD. students expose and monetize APIs across IBM enabled platforms

iCEOD Staff, JKUAT faculty, PhD students and other Research Partners enrich and extend the Open Data by applying analytics to unlock valuable insights

Developers, ISVs, Entrepreneurs access Open Data through a catalogue of free and monetized APIs using these to create innovative apps and solutions

End Users take advantage of these innovative Apps to address their needs and problems, commercialised paid and free Apps with in-app purchases

CODATA Recommended Values of the Fundamental Physical Constants, 2014: http://dx.doi.org/10.5281/zenodo.22826

2014 CODATA RECOMMENDED VALUES OF THE FUNDAMENTAL CONSTANTS OF PHYSICS AND CHEMISTRY NIST SP 959 (Aug 2015)

See: P. J. Mohr, D. B. Newell, and B. N. Taylor, arxiv.org/pdf/1507.07956v1.pdf (2015). A more extensive listing of constants is available in the reference given above and on the NIST Physical Measurement Laboratory Web site: physics.nist.gov/constants.

Quantity	Symbol	Numerical value	Unit
speed of light in vacuum	c, c_0	299 792 458 (exact)	$\mathrm{m\ s^{-1}}$
magnetic constant	μ_0	$4\pi \times 10^{-7}$ (exact)	$\mathrm{N\ A^{-2}}$
electric constant $1/\mu_0 c^2$	ϵ_0	$8.854\,187\,817... \times 10^{-12}$	$\mathrm{F\ m^{-1}}$
Newtonian constant of gravitation	G	$6.674\,08(31) \times 10^{-11}$	$\mathrm{m^3\ kg^{-1}\ s^{-2}}$
Planck constant	h	$6.626\,070\,040(81) \times 10^{-34}$	J s
$h/2\pi$	$\hbar$	$1.054\,571\,800(13) \times 10^{-34}$	J s
elementary charge	e	$1.602\,176\,6208(98) \times 10^{-19}$	C
fine-structure constant $e^2/4\pi\epsilon_0\hbar c$	α	$7.297\,352\,5664(17) \times 10^{-3}$	
inverse fine-structure constant	α^{-1}	137.035 999 139(31)	
Rydberg constant $\alpha^2 m_e c/2h$	R_∞	10 973 731.568 508(65)	$\mathrm{m^{-1}}$
Bohr radius $\alpha/4\pi R_\infty$	a_0	$0.529\,177\,210\,67(12) \times 10^{-10}$	m
Bohr magneton $e\hbar/2m_e$	μ_B	$927.400\,9994(57) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$

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

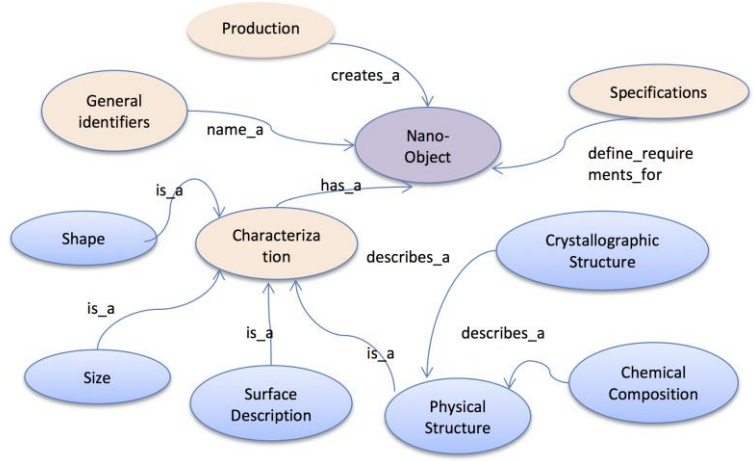


Figure 4. Information categories for describing an individual nano-object

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

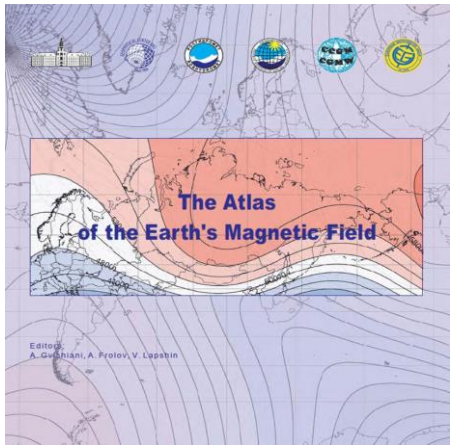
Uniform Description System v.02, May 2016:
http://dx.doi.org/10.5281/zenodo.56720

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

Challenges in Data Science

TG Earth-Space Science Data Interoperability

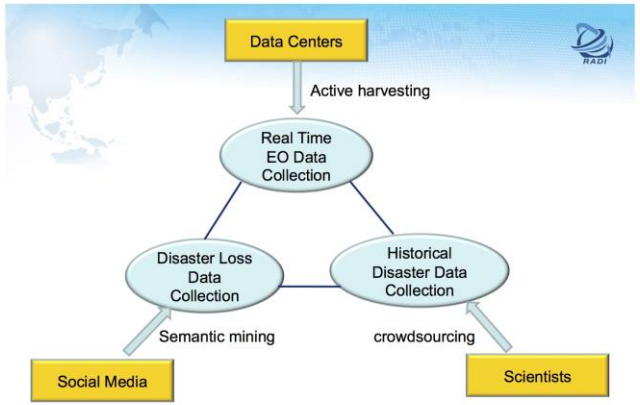
Preparing second edition of Atlas of the Earth's Magnetic Field http://bit.ly/atlas-magnetic-field



- Contribution to standards for multidisciplinary GIS for geoscience data
- **Increased focus on interoperability and standardisation issues.**
- International collaboration for conferences and training activities (Moscow and Sochi, July 2016; Peterhof, October 2017).

TG LOD Global Disaster Risk Research

- White Paper: 'Gap Analysis on Open Data Interconnectivity for Global Disaster Risk Research' http://bit.ly/White_Paper-LOD_Disaster_Gap_Analysis
- Important response from the perspective of science and data to post-Sendai framework
- **Inviting comments until 30 September.**



Challenges in Data Science

TG Coordinating Data Standards amongst Scientific Unions

- Encourage increased coordination and collaboration on vocabularies and ontologies across International Scientific Unions.
- Compile and distribute information about ISU data and information standards.
- Develop a maturity model and good practice guidelines.
- Identify opportunities for increased technical and semantic coordination.

TG Agricultural Data for Knowledge and Innovation

- Coordinating development of policy, more effective application of standards and capacity building/training.
- Initial focus on EAR (East African Region).

IRIDIUM International Glossary of Research Data Terms

- Building on a glossary first developed by Research Data Canada.
- CODATA helping to internationalise the glossary.
- Adopting CASRAI processes for review and updating in systematic working group and review circle cycles.
- Current Glossary: http://dictionary.casrai.org/Category:Research_Data_Domain
- Information about the 2017 process: http://bit.ly/IRIDIUM-RDM-Glossary

casrai
Connecting Research

RDC DRC
Research Data Canada – Données de Recherche Canada

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 Research Data Science 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

CODATA-RDA School of Research Data Science

- **First School of Research Data Science, 1-12 August 2016, ICTP, Trieste**
- Funding for students and tutors provided by ICTP, TWAS, CODATA, ACU, RDA Europe, GEO and GODAN.
- Attended by 70 students from all around the world.

The Abdus Salam International Centre for Theoretical Physics

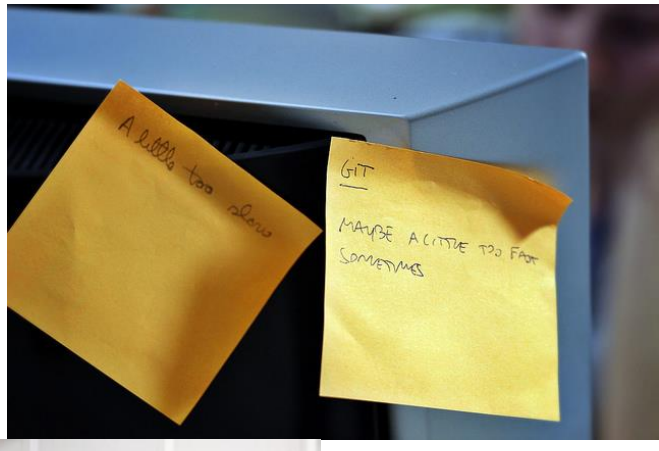
RESEARCH DATA ALLIANCE EUROPE

The Association of Commonwealth Universities

GODAN Global Open Data for Agriculture & Nutrition

#DataTrieste

#DataFoo…



- **Programme for #datatrieste** http://bit.ly/School_of_Research_Data_Science-Programme
- **School will repeat at Trieste in 2017 and 2018, *at least…***
- Possibly with addition of one week more advanced on Big Data.
- **Will run foundational two week course at ICTP INESP in Sao Paolo, Brazil, December 2017.**
- Schools can be run with a greater or lesser degree of support and coordination from the international convenors.
- **Keen to encourage a network of schools, but also local schools with lower central input.**

- Discussions with possible partners in South Africa and India.
- Keen to explore opportunities with CODATA National and Union Members.

INTERNATIONAL COUNCIL FOR SCIENCE

ICSU CODATA

Thank you for your attention!

Credits for slides: inc. Geoffrey Boulton, Joseph Muliaro Wafula
Credit for photos: Andjani Gatzweiler

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