Open Data and the Developing World
GEO-XI Plenary, DSWG Side Event, WMO, Geneva
Wednesday 12 November 2014

CODATA Capacity Building and the Data Sharing Principles in Developing Countries

Simon Hodson
Executive Director CODATA
www.codata.org/blog
execdir@codata.org
@simonhodson99

CODATA: international collaboration for Open Scientific Data

Data Policy

- International and national aspects of data policy.
- Data policy committee: setting an international agenda for data policy, expert forum, advice and consultancy.
- Coordinating with national committees.

Data Science

- Long-standing activities: fundamental constants.
- Strategic working groups; community-driven task groups.
- Disciplinary and interdisciplinary data challenges, Big Data

Capacity Building

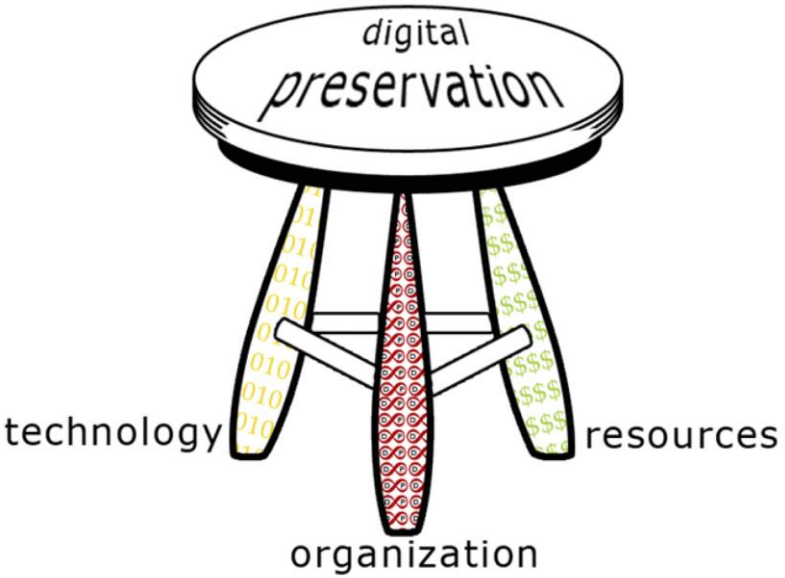
- Longstanding work on data preservation and access with developing countries.
- Executive Committee Task Force on Capacity Building: setting an international agenda for capacity building; Early Career WG

Data for International Science

- Support for ICSU Mission.
- Data issues and challenges in international, interdisciplinary science programmes.

Why do policies matter?

- A necessary first step: necessary but insufficient.
- Important to set a policy framework.
- Three-legged stool model for institutional (or national) capacity for digital preservation.
- Developed at Cornell and ICPSR.
- Informs maturity models developed by UK Digital Curation Centre.
- Organisational infrastructure: policies, processes, practices.
- Technological infrastructure: kit and standards.
- Resource infrastructure: sustainability.



PASTD in Developing Countries TG

- CODATA Task Groups offer seed money only, helps coordinate and target activity.
- CODATA brings to the table a network of motivated and expert data scientists.
- Series of workshops on data access and preservation issues in developing countries.
  - Asia: Beijing, China 2003, 2004, 2006; Shanghai, China 2008; Tibet, 2012; Ulaanbaatar, Mongolia (2008, 2010); Malaysia (2015 in planning)
  - Africa: Pretoria, South Africa (2002, 2007); **Nairobi, Kenya (2014)**
  - South America: Sao Paulo, Brazil (2007); Havana, Cuba (2009); Bogota, Colombia (2013).
- Partnerships with international organisations: IAP (International Academy Panel), Asia Science Council, APN (Asia Pacific Network for Global Change Studies), UNESCO.
- CODATA partnered in IAP *International Issues Program on Digital Knowledge Resources and Infrastructure in Developing Countries*. Developed a 'toolkit' for setting up Institutional Repositories.
- Importance of partnership with local institutions and organisations.
- Practical methodology, focusses on **open data as a common necessary factor in scientific discovery and scientific application** (key elements of the knowledge economy).
- **Training workshop focuses on awareness raising, general data management and archiving principles, specific technical aspects (e.g. use of GIS).**

CODATA Strategic Commitment

- Executive Committee led activities to build on and support PASTD and ECDP.
- Set a **Smart Agenda for Capacity Building** in data management and data science
- **Definition** of Capacity Building
- **Horizon Scan** based on survey of CODATA members, aim identify specific needs, good resources, connect people for knowledge exchange and training.
- **Curriculum Framework**
- Clearing house for **training materials** - with particular attention to those of relevance for developing countries
- **Advanced workshop and data ambassador programme**
- **Coordinate activities/proposals to obtain funds!**

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*** and ***Guidelines*** for their implementation.
- Particular concerns for credit and periods of privileged access; concerns about giving away intellectual assets for free…

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

1. Data should be open and unrestricted.
2. Data should be free to the user.
3. Data should be informative and assessed for quality.
4. Data sharing should be timely.
5. Data should be easy to find and access.
6. Data should be interoperable
7. Data should be sustainable.
8. Data contributors should be given credit.
9. Data access should be equitable.
10. Data may be restricted, in exceptional cases, if adequately justified.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

4) Data sharing should be timely.
Once datasets are sufficiently informative and quality controlled, they should be released as quickly as possible. This can be done in steps, starting with the metadata to avoid duplication. **In some cases, such as public emergencies and disasters, open release of relevant data should be an immediate priority. In other cases, such as research, data should be openly available no later than upon the publication or patenting of results.** Users in developing countries have the most to gain from such policies.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

8) Data contributors should be given credit.
A significant incentive for the open disclosure and "publication" of a dataset is the ability to properly cite and attribute the contributor(s), whether internal or external to an organization. Any subsequent user of the data has at least an ethical obligation—and possibly a legal one—to cite and attribute the source of the data whenever they are reused, and not to misuse the data in any way. Such practices can also improve the integrity of the data sets made available by the contributors, in support of Principle 3. In particular, **data contributors in the developing world require recognition and rewards for such disclosure, and this should become common practice.** A persistent digital identifier, attached to the dataset online, is the best way to promote this goal.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

9) Data access should be equitable.
Open access and use of data in developing countries, especially for public purposes, should be supported by the governments and institutions in the more economically developed nations. **Capacity building of essential experts and infrastructure in developing countries should be a priority of international organizations.** Similarly, experts in developing countries should join and actively participate in the relevant regional and international organizations.

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.
- **Funding Proposal for planning activity and first workshop with Jomo Kenyatta University of A&T and Kenyan National Academy of Sciences**
- **'Promote the enhanced data skills for science which are needed for Kenyan scientists to make better use of Kenya's own data resources.'**

CODATA-ICSTI Task Group
Data Citation, Standards and Practices

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

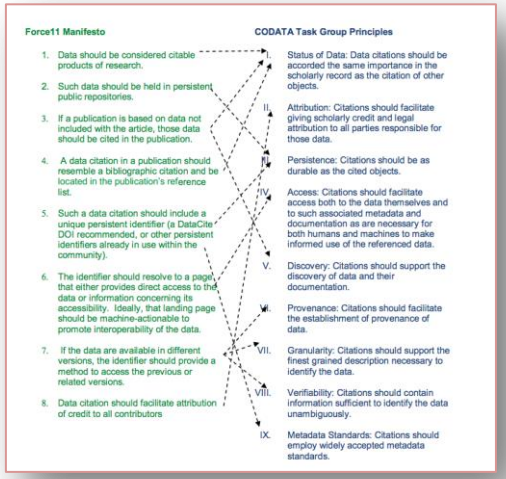
Out of Cite, Out of Mind
http://bit.ly/out_of_cite

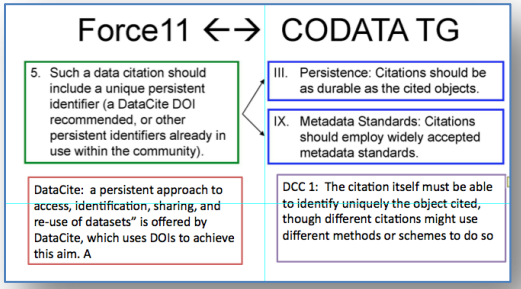
Joint Declaration of Data Citation Principles:
https://www.force11.org/datacitation

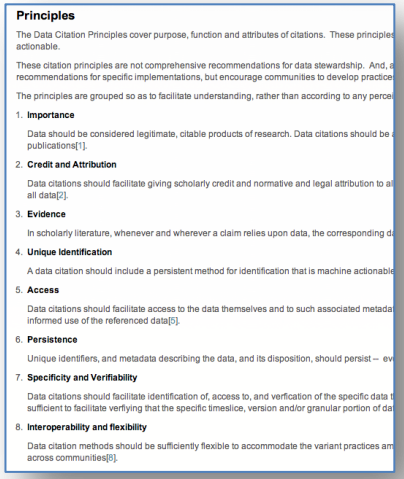
Background and Developments:
http://bit.ly/data_citation_principles

DC1
Data Citation Principles

Process







Synthesis → Community feedback → Revision → Dissemination

July-Sept 2013 | Nov-Dec 2013 | Jan 2014 | Now

Data Citation Principles: Open for Endorsement

Joint Declaration of Data Citation Principles (Overview); Slide: Sarah Callaghan

Joint Data Citation Principles

Purpose

1. Importance. Data should be considered legitimate, citable products of research. **Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.**

2. Credit and attribution. Data citations should facilitate giving scholarly **credit** and normative and legal **attribution** to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

3. Evidence. In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.

Function

4. Unique Identification. A data citation should include a persistent method for identification that is machine-actionable, globally unique, and widely used by a community.

5. Access. Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.

Joint Data Citation Principles

Attributes

6. Persistence. Unique identifiers, and metadata describing the data and its disposition, should persist -- even beyond the lifespan of the data they describe.

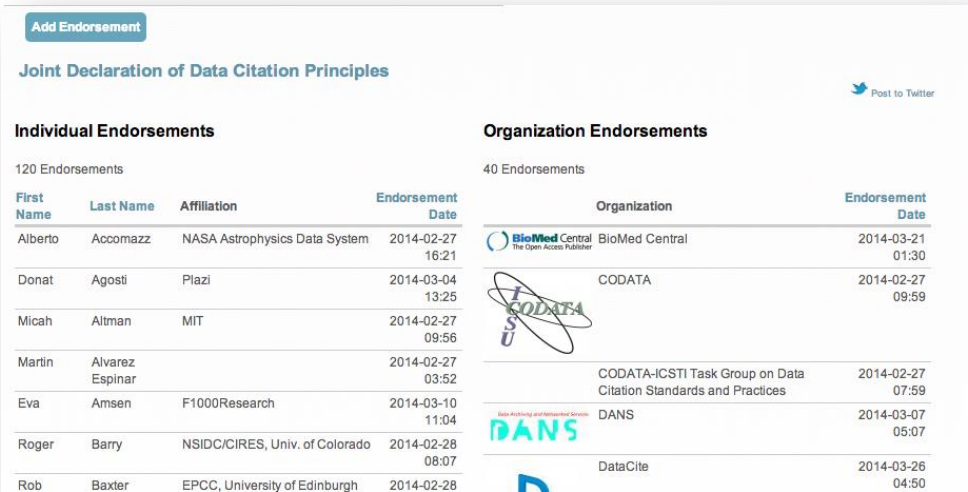
7. Specificity and verifiability. Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

8. Interoperability and flexibility. Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities [8].

Endorse the Data Citation Principles

https://www.force11.org/datacitation/endorsements

FORCE11
The Future of Research Communications and e-Scholarship



Joint Declaration of Data Citation Principles (Overview); Slide: Sarah Callaghan

Thank you!

CODATA Website: http://www.codata.org/
CODATA Blog: http://www.codata.org/blog/
SciDataCon 2014: http://www.scidatacon2014.org/
CODATA General Assembly 2014: http://www.codata.org/general-assembly

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

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

1) **Data should be open and unrestricted.** Data generated with public support, including private foundations, should be openly accessible and subject to unrestricted (re)use, absent specific, justified reasons to the contrary (see Principle 10). Openness is especially beneficial for development purposes and research uses, but can benefit all society equally and have a multiplier effect on the economy.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

2) Data should be free to the user. In most cases, any cost for access is an insurmountable barrier to users in the developing world. Therefore, data should be free online to the user. In some special cases, access to data may be no more than the marginal cost of filling a user request. At the same time, it is recognized that adequate preparation and open availability of data require sufficient financial support (see Principle 7).

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

3) Data should be informative and assessed for quality. Data should be of known quality and integrity, and should be organized and described (with metadata) in datasets sufficient to allow them to be understood and effectively (re)used by others. Baseline technical and management standards need to be established, especially in the developing world where state-of-the art practices are not yet as prevalent. Adequate preparation and the use of non-proprietary software are especially important for any datasets expected to have long-term value.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

4) Data sharing should be timely.

Once datasets are sufficiently informative and quality controlled, they should be released as quickly as possible. This can be done in steps, starting with the metadata to avoid duplication. In some cases, such as public emergencies and disasters, open release of relevant data should be an immediate priority. In other cases, such as research, data should be openly available no later than upon the publication or patenting of results. Users in developing countries have the most to gain from such policies.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

5) Data should be easy to find and access. Upon the public release of any dataset, the provider should promote ease of access by the broadest user base. Diverse means of publication should be considered in recognition of potential connectivity and other technological challenges.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

6) Data should be interoperable, when necessary. If data from a dataset are likely to be combined with data from one or more other datasets (e.g., in geospatially referenced research), special attention should be given to making such data technically, semantically, and legally interoperable.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

7) Data should be sustainable. The life-cycles of any datasets that are expected to be reused by others should be planned at the outset with support sufficient to successfully implement the first six Principles. The lower availability of funding in developing countries, especially for long-term preservation, makes this a key priority so that valuable datasets remain intelligible and are not lost or in need of rescue. Cost recovery for data archiving and availability should not be borne by the users, consistent with Principle 2, but by other entities in the data lifecycle.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

8) Data contributors should be given credit. A significant incentive for the open disclosure and "publication" of a dataset is the ability to properly cite and attribute the contributor(s), whether internal or external to an organization. Any subsequent user of the data has at least an ethical obligation—and possibly a legal one—to cite and attribute the source of the data whenever they are reused, and not to misuse the data in any way.

Such practices can also improve the integrity of the data sets made available by the contributors, in support of Principle 3. In particular, data contributors in the developing world require recognition and rewards for such disclosure, and this should become common practice. A persistent digital identifier, attached to the dataset online, is the best way to promote this goal.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

9) Data access should be equitable.

Open access and use of data in developing countries, especially for public purposes, should be supported by the governments and institutions in the more economically developed nations. Capacity building of essential experts and infrastructure in developing countries should be a priority of international organizations. Similarly, experts in developing countries should join and actively participate in the relevant regional and international organizations.

Data Sharing Principles in Developing Countries Developed at Nairobi Workshop

10) Data may be restricted for a limited time, if adequately justified. Restrictions may be placed on access to and uses of publicly funded data and datasets for specified periods of time. Justified restrictions may include specific protections of national security, personal privacy, intellectual property, confidentiality, and other values, such as indigenous peoples' rights or location of endangered species. Nevertheless, the default rule should be one of openness, consistent with Principle 1, and any restrictions should be minimized to the extent possible.