Data Citation (Out of cite, out of mind: Current state of play)

Presentation by Martie van Deventer
to
eResearch Africa 2013 Conference
08 October

With slides by:
Paul F. Uhlir,
Board on Research Data and Information (BRDI), National Research Council, National Academy of Sciences

Daniel Cohen,
Board on Research Data and Information (BRDI), National Research Council, National Academy of Sciences (on detail from Library of Congress)

Sarah Callaghan,
British Atmospheric Data Centre, UK

Roadmap

- Background about the CODATA-ICSTI data citation project
- Task group work:
  - Literature survey
  - Core elements of a citation
  - Data citation examples
  - Citation practice in South Africa (very limited sample)
- Out of cite report
- What to do?

Organizations Investigating Data Citation

- **International Council for Scientific and Technical Information (ICSTI) through CODATA**
- **DataCite**
- The Dataverse Network
- National Information Standards Organization (NISO)
- Creative Commons
- CENDI – U.S.
- Global Biodiversity Information Facility (GBIF)
- World Data System (WDS)
- STM-Association
- Digital Curation Center, UK
- **Research Data Alliance (RDA)**

Managing Organizations for our research

- International CODATA Task Group on Data Citation Standards and Practices
  http://www.codata.org/taskgroups/TGdatacitation/index.html
  Approved at CODATA 27th General Assembly in Cape Town, SA 2010
- BRDI
  http://www.nas.edu/brdi
  Ad hoc committee of the **Board on Research Data and Information**, at the **U.S. National Academy of Sciences**, in Washington, DC. BRDI represents the U.S. National Committee for CODATA.
- BRDI staff supports both projects.

FUNDING

We are grateful to the following funders of this project:

- CODATA
- Institute for Museum and Library Services
- Library of Congress
- Microsoft Research
- Sloan Foundation

ICSTI-CODATA Data Citation Task Group

Co-Chairs:

Jan Brase,(Director, DataCite, and ICSTI representative), Technische Informations Bibliothek (TIB)/German National Library of Science and Technology, GERMANY

Sarah Callaghan (U.K. CODATA), The NCAS British Atmospheric Data Centre, STFC Rutherford Appleton Laboratory, UNITED KINGDOM

Bonnie Carroll (U.S. CODATA and CENDI), President, Information International Associates, USA - till Jan 2013

Christine Borgman, University of California, Los Angeles, USA – as of Jan 2013

Members:

Micah Altman, USA

Elizabeth Arnaud, ITALY

Christine Borgman, USA

Todd Carpenter, USA

Dora Ann Lange Canhos, BRAZIL

Vishwas Chavan, DENMARK

Nathan Cunningham, UNITED KINGDOM

Michael Diepenbroek, GERMANY

Puneet Kishor, USA

Mark Hahnel, UNITED KINGDOM

John Helly, USA

Jianhui LI, CHINA

Franciel Azpurua Linares, USA

Brian McMahon, UNITED KINGDOM

Karen Morgenroth, CANADA

Yasuhiro Murayama, JAPAN

Fiona Murphy, UNITED LINGDOM

Giri Palanisami, USA

Mark Parsons, USA

Soren Roug, BELGIUM

Helge Sagen, NORWAY

Eefke Smit, THE NETHERLANDS

Martie J. van Deventer, SOUTH AFRICA

John Wilbanks, USA

Michael Witt, USA

Koji Zettsu, JAPAN

Consultants:

Daniel Cohen, Library of Congress ,USA

Franciel Linares, Information International Associates, USA

Yvonne Socha, MLIS candidate, University of Tennessee, USA

Paul F. Uhlir, U.S. National Committee for CODATA and Board on Research Data and Information, National Academy of Sciences, USA

Task Group Objectives and Deliverables

- Conduct inventory and analysis of existing **literature** and existing data citation and attribution **initiatives**.
- Investigate and analyze **how existing data repositories cite** and provide attribution to their data sets.
- Identify and obtain **input from stakeholders** in the library, academic, publishing and research communities.
- Provide an **international forum** to identify and help reconcile the needs of various stakeholder communities.
- Share information and create greater awareness of these issues internationally.
- Establish a **public web presence**.
- Conduct meetings and workshops to articulate the state of the art and best practices in this area, and to identify emerging issues.
- Work with the major international, regional, and national standards organizations to **develop formal data citation and attribution standards and best practices**.
- **Promote scientific data attribution** by developing models, tools, and practical guidance on how to publish citable and trackable data sets.

Schedule of Activities

Completed, ongoing

- Bibliographic inventory and analysis (literature review) (ongoing).
- Symposium and workshop held in Berkeley, CA in August 2011.
- Interviews with a sample of identified stakeholders concerning data citation and attribution practices
  - Data Repositories
  - Publishers
  - Researchers
  - Funding Organizations
- Publish Report from August '11 Symposium and Workshop (Jun 2012).
- Out of cite out of mind: The current state of practice, policy and technology— September 2013.

… and planned

- Active dissemination of first phase results in 2012-2013. Examples:
  - Sponsored Session at CODATA International Conference in Taipei, TW October/November 2012
  - STM Innovations Seminar. April 30, 2013
  - 5th African Conference for Digital Scholarship and Curation. Durban, Jun 2013
  - eResearch Africa 2013. Cape Town. October 2013
- Principles and Best Practices White Paper Workshop in September 2013
- White Paper disseminated 2013 - 2014.

Literature review

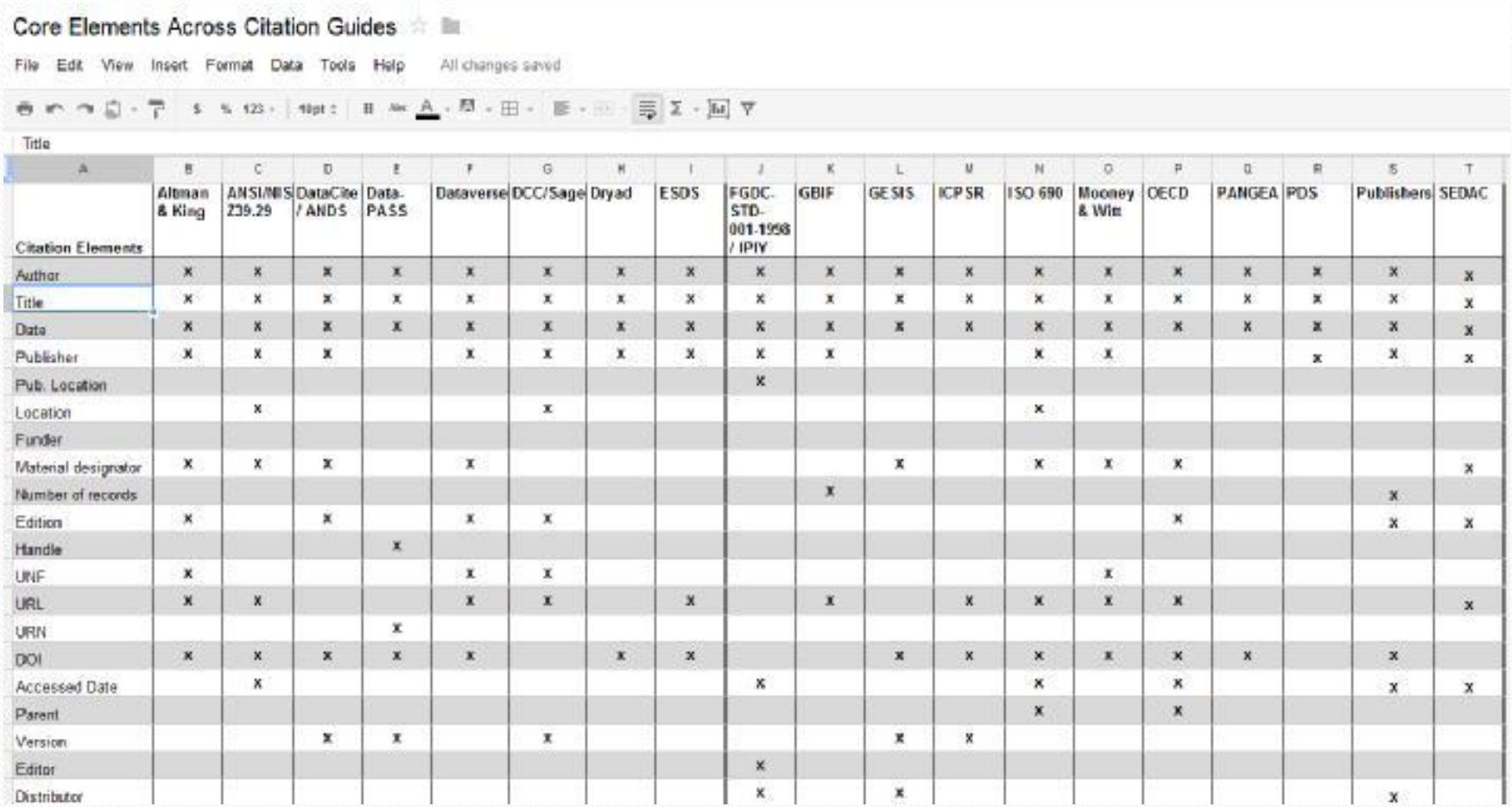
- 384 resources in 15 different formats (& growing)
- Mainly research papers
- Facets addressed: policies, infrastructure, research practices, and best practices development
- Also: **Linked data**, dynamic data, open data, data set management practices (general or for different scientific fields such as biology), technology such as infrastructure & system architecture, **unique identifiers**, semantic web, digital data collection, attribution, contributor identifier, dissemination, collaboration and sharing, **preservation**, archival, verification, provenance, the use of ontologies, repositories, data usage & metrics, data publishing, geospatial data management

Elements of a data citation

- **Author**
- **Title**
- **Date** (of publication)
- **Publisher**
- **URL/ URI/ UNF /** (electronic retrieval locator)
- **Persistent identifier** (DOI/ Handle)
- **Resource type**
- **Location**
- **Version**
- **Funder**
- **Material designator**
- **Edition**
- **Accessed date**
- **Parent series**
- **Accession number**
- **Notes**

Ball & Duke (2011), Mooney & Newton (2012) and others

Core citation elements



Core Elements Across Citation Guides

File Edit View Insert Format Data Tools Help All changes saved

Title

Citation Elements	Altman & King	ANSI/NISO Z39.29	DataCite / ANDS	Data-PASS	Dataverse	DCC/Sage	Dryad	ESDS	FGDC-STD-001-1998 / IPIY	GBIF	GESIS	ICPSR	ISO 690	Mooney & Wim	OECD	PANGEA	PDS	Publishers	SEDAC
Author	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Title	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Date	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Publisher	x	x	x		x	x	x	x	x	x			x	x			x	x	x
Pub. Location									x										
Location		x				x							x						
Funder																			
Material designator	x	x	x		x						x		x	x	x				x
Number of records										x								x	
Edition	x		x		x	x									x			x	x
Handle				x															
UNF	x				x	x								x					
URL	x	x			x	x		x		x		x	x	x	x				x
URN				x															
DOI	x	x	x	x	x		x	x			x	x	x	x	x	x		x	
Accessed Date		x							x				x		x			x	x
Parent													x		x				
Version			x	x		x					x	x							
Editor									x										
Distributor									x		x							x	

Data citation example: DCC

Cool, H. E. M., and Bell, M. 2001. "Excavations at St Peter's Church, Barton-upon-Humber." Archaeology Data Service. Accessed: 1 May 2011. http://dx.doi.org/10.5284/1000389

Data Citation Example – ESIP Federation

D. Cline, R. Armstrong, R. Davis, K. Elder, and G. Liston. 2002, Updated 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. Parsons and M. J. Brodzik. National Snow and Ice Data Center. Data set accessed 2008-05-14 at http://dx.doi.org/10.5060/D4MW2F23z

Requirements for data citation

- The Citation should
  - uniquely identify the object cited.
  - support the retrieval of the cited object.
  - be human readable.
  - be machine ‘processable’.
- The citation mechanism should be compatible with Web infrastructure.
- The citation ‘system’ should be able to generate a citation with all the desired fields.
- The citation mechanism should be identifier-agnostic.

Stakeholder consultation: Berkeley, CA August 2011

- Data centers/ repositories
- Research funders
- Researchers (also though professional societies)
- Publishers and editors

Also conducted country specific interviews Jan – April 2012

Symposium and Workshop Sessions

I. Why are attribution and citation of data **important**?

II. Major **technical issues** in developing and implementing scientific data citation standards and practices

III. Major **scientific issues** in developing and implementing scientific data citation standards and practices

IV. Major **institutional, financial, legal, and socio-cultural issues** in developing and implementing scientific data citation standards and practices

V. **Status of data attribution and citation practices** in the natural and social sciences in the U.S. and internationally

VI. **Institutional roles and perspectives**: similarities and differences across disciplines and countries

VII. Workshop – Options: Where do we go from here?

For more information on the symposium and workshop outcome see:
http://sites.nationalacademies.org/PGA/brdi/PGA_063656

Country specific research: Questions to researchers

- Do you share data outside of your collaborations?
- Have your data ever been cited?
- Have you had any particular problems that hindered citation of your data?
- Have you had any particular problems that hindered **you** from citing data?

Excising citation practice: SA

- Very small sample!
- There is no standard practice but the research discipline appears to determine behaviour.
- Usually share their data when personal requests are received. Confidentiality often causes a barrier in sharing.
- Spatial data researchers appear more aware & often share/ use data sets. Others have not given the sharing of data much thought.
- Majority not aware that their data is being cited. They see citation as a sign of courtesy rather than mandatory.
- Majority not aware of existing citation standards & guidelines.
- Not interested in re-processing the data so that others could make use it.
- Some have used and cited data but the majority have not.

Outline of current state of practice report

1. Importance of data citation
2. Defining the concepts
3. Emerging principles for data citation
4. Institutional infrastructure for data citation
5. Technical infrastructure
6. Benefits and challenges for good citation practices
7. Open research questions

Was released end September 2013

CODATA-ICSTI. 2013. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal.* V12. Available: https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article

Importance of data citation

- Scientific tradition to share findings.
- Citing the use of the findings recognizes/ acknowledges the contribution.
- In the past data was recorded as graphs, tables & images in an article.
- But … data is a cornerstone of research & data sets have become useful research outputs.
- Funders require data deposit or at least encourage the deposit of data in repositories - discoverability.

Defining the concepts

- Application to general, archival and infrastructural contexts – to get to common jargon.
  - Data objects
  - Data preservation
  - Citation and metadata
- Also investigated distinctions among:
  - Literature-to-data,
  - Data-to-literature, and
  - Data-to-data citations.

Emerging principles for data citation

- 10 principles to guide the ‘what to do’ practice
  - Data is of equal importance as other scientific objects
  - Citations should facilitate giving credit to responsible parties
  - Citations should be as durable as the object cited
  - Access to data is necessary for both humans & machines
  - Citations should support discoverability of data & related documents
  - Citations should facilitate the provenance of the data
  - Provide the finest grain description of the data
  - Identify the data unambiguously
  - Employ widely acceptable metadata standards
  - Citation practices should support interoperability across data communities

Institutional infrastructure for data citation

- Identified the key players in establishing citation standard.
  - International scientific organizations
  - Researchers & research institutions
  - Publishers & scholarly journals
  - Academic & research libraries
  - Research funding agencies
- Surveyed current policy & operational experience – various disciplines, various organizations.

Technical infrastructure

- It is not difficult technically to develop data citation protocols.
- Granularity, versions, micro-attribution, contributor identifiers, and derivatives are issues for debate.
- Mainly socio-cultural, institutional and economic barriers that prohibit the uptake of technical protocols.
- Deeper understanding, of how technologies facilitate the use and re-use of data, needs to be developed.

Benefits and challenges for good citation practices

- Listed benefits & challenges – for a wide variety of stakeholders (researcher to funder).
- Suggested ways & means to overcome especially socio-cultural and institutional challenges.

Open research questions

- Research required to guide a maximally effective data citation system.
- Types of new metrics & domain research that data citation aims to enable.

What next?

- Data Citation Principles and Best Practices White Paper workshop took place in September 2013 (Washington).
- Dissemination of the White Paper planned for distribution in 2014

In the absence of a standard - Advice to researchers

- Deposit your data in a trusted repository – where a persistent identifier is allocated.
- Cite datasets that you make use of.
- Provide dataset identifiers in the form of a URL/ DOI/ Handle wherever possible.
- Include data citations alongside citations for textual publications.
- Cite datasets at the finest-grained level possible.
- Make sure that you cite the exact version of the data used.
- Notify the repository when using one of their datasets – so that they could create links also to your paper.

Ball & Duke (2011)

In the mean time …

- Suggest that you request advice from your citation style producer.
- Make use of the DCC citation style guide by Ball & Duke (rev 2012).
- Most important: ensure that your data is made available and that you provide the preferred citation style to potential users of your data.

Questions?

Further information:

mvandeve@csir.co.za, sarah.callaghan@stfc.ac.uk , puhlir@nas.edu or dcohen@nas.edu

References

- Ball, A. & Duke, M. rev 2012. How to cite datasets and link to publications. Available: http://www.dcc.ac.uk/resources/how-guides/cite-datasets
- Borgman, C.L. 2010. "Research Data: Who will share what, with whom, when, and why?" China-North America Library Conference, Beijing
  Available: http://works.bepress.com/borgman/238
- Mooney, H., Newton, M.P. 2012, The anatomy of a data citation: Discovery, reuse, and credit. Journal of librarianship and scholarly communication. V1(1). Available: http://jlsc-pub.org/cgi/viewcontent.cgi?article=1035&context=jlsc
- CODATA-ICSTI. 2013. Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal.* V12. Available: https://www.jstage.jst.go.jp/article/dsj/12/0/12_OSOM13-043/_article
- Callaghan, S. 2012. Data citation standards and practices. Presentation to DataCite summer meeting Copenhagen, June 14th, 2012
- Uhlir, P.E. (Rapporteur). 2012. For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop (Berkeley, CA on August 22-23, 2011). Available: http://www.nap.edu/catalog.php?record_id=13564