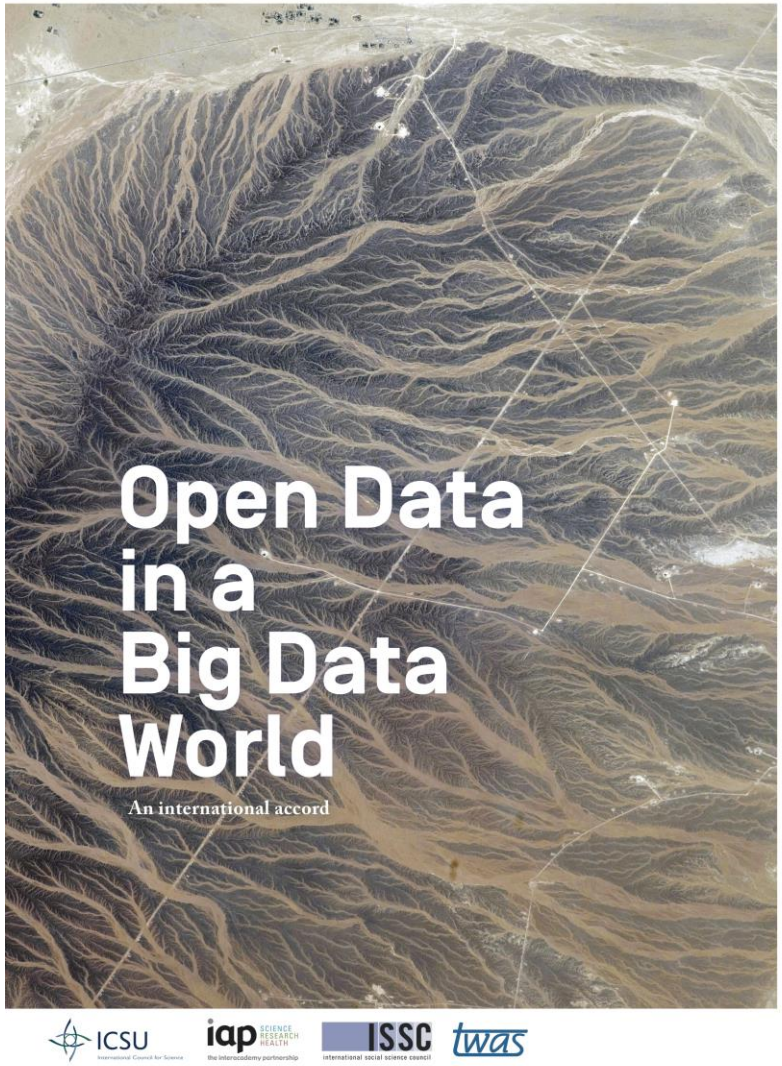
Data citation as a catalyst for good RDM practices
CSIR, Pretoria, South Africa
10 December 2015

Data Citation: framing the discussion and global context

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

The Case for Open Data in a Big Data World

- **Science International Accord on Open Data in a Big Data World:** http://bit.ly/opendata-bigdata
- Presents a powerful case that the profound transformations mean that data should be:
  - Open by default
  - Intelligently open
- Supported by four major international science organisations.
- Lays out a framework of principles for how the vision of Open Data in a Big Data World can be achieved.
- Parallel proposal for an African open science capacity initiative.



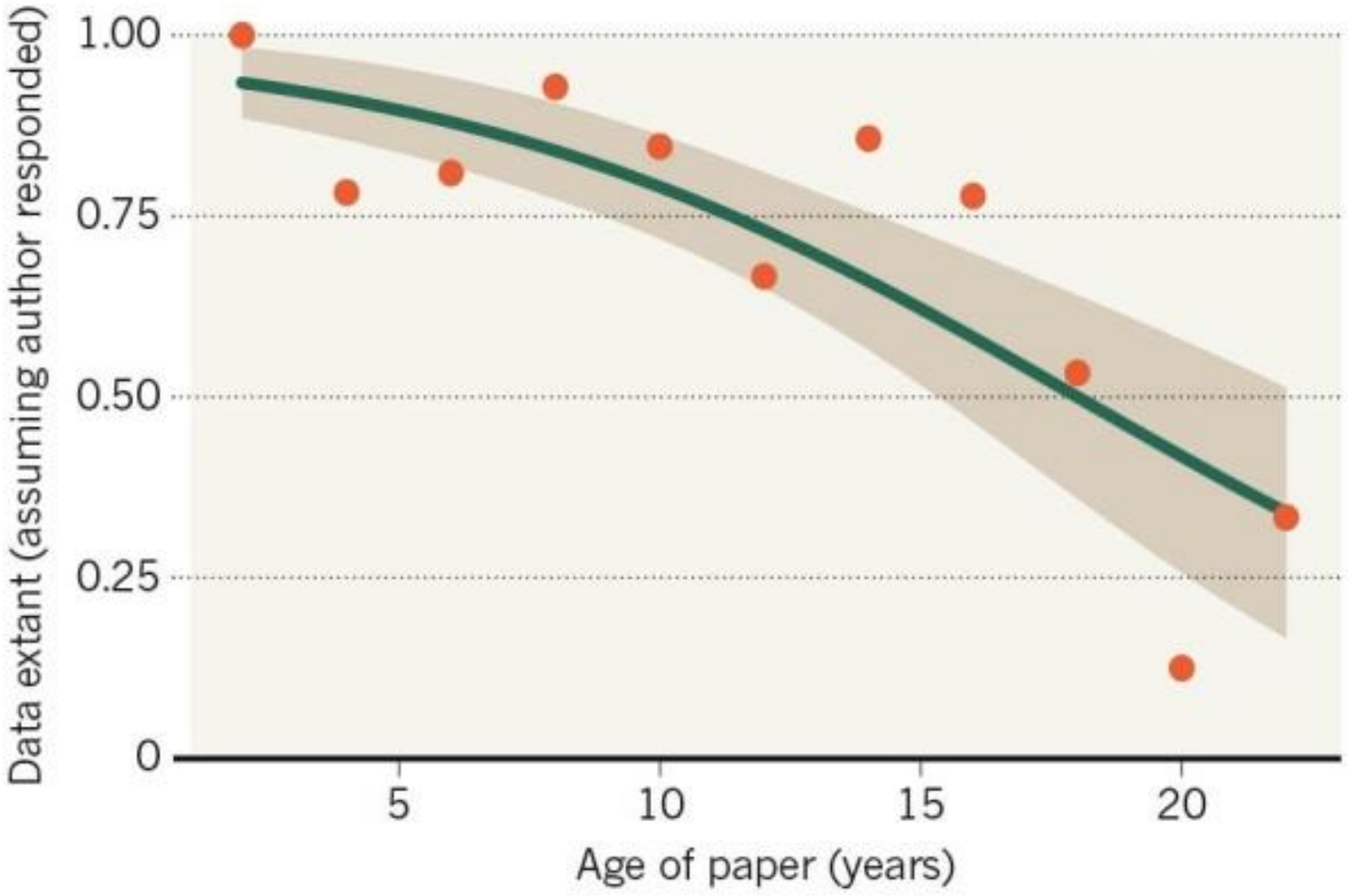
Benefits of Open Data: some examples from GEO

- **Barbara Ryan, Director of Secretariat GEO, TED-X Talk Barcelona**
- In 2008 US Government was convinced to make Landsat Data openly available, for free.
- Under charging, the highest number of downloads was 53 scenes per day.
- Now over 5700 scenes per day are downloaded.
- Spanish deforestation research: under the charging regime data access alone would have cost €260M
- **CODATA produced a White Paper on the Value of Data Sharing for the GEO-XII Plenary: http://dx.doi.org/10.5281/zenodo.33830**

https://www.youtube.com/watch?v=9umWTFgFIVs

GROUP ON EARTH OBSERVATIONS

80% of ecology data irretrievable after 20 years



Vines TH *et al*. (2013) Current Biology DOI:10.1016/j.cub.2013.11.014

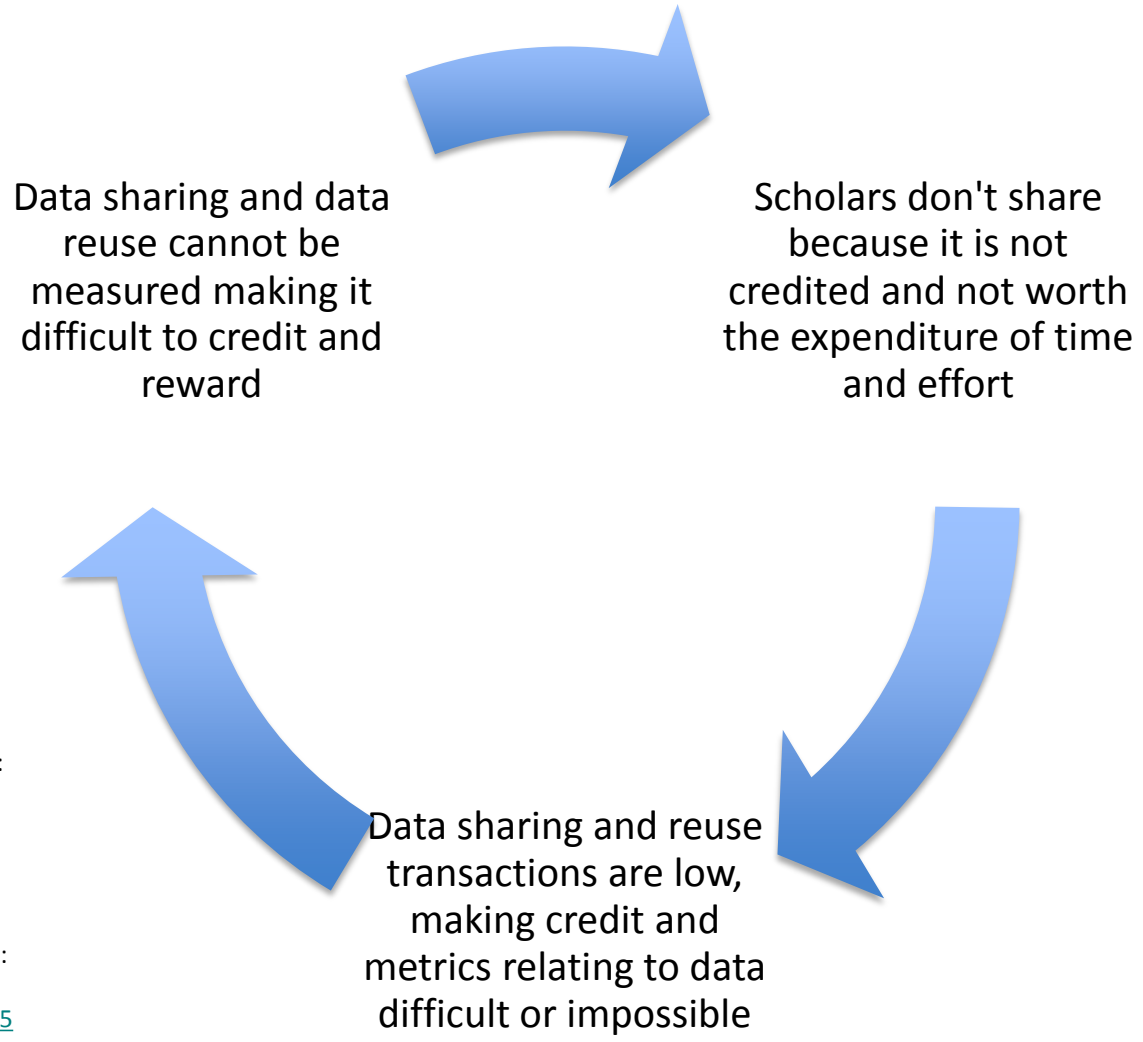
Barriers to Data Sharing

Researchers concerns:

- Concern that data may be misused or misunderstood.
- Concern that will lose scientific edge if sharing before fully exploited.
- Desire to retain control of a professional asset.
- **Concern that will not be credited.**
- **Lack of career rewards for data publication.**
- See ODE report, using Parse.Insight findings: http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf
- Culture in particular research disciplines; availability of infrastructure.
- **Fundamentally, researchers are reluctant to expend effort sharing data because they do not feel that data is adequately exposed or credited.**

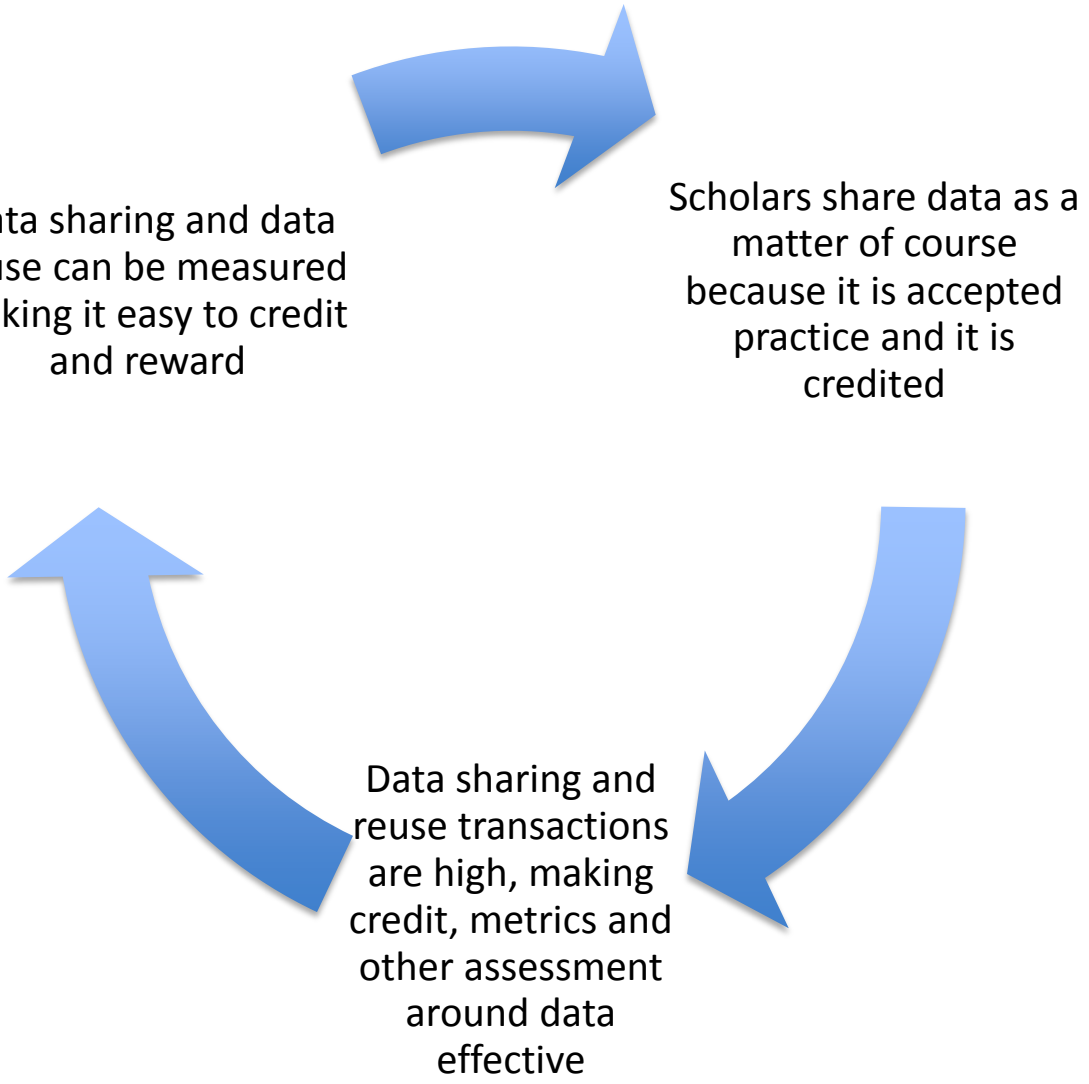
Nature special issue on data sharing: http://www.nature.com/news/specials/datasharing/index.html

Credit for Data: from a **vicious** to a virtuous circle?



See The Value of Research Data: Metrics for datasets from a cultural and technical point of view http://www.knowledge-exchange.info/datametrics and Knowledge Exchange Workshop: 'Making Data Count': http://www.knowledge-exchange.info/Default.aspx?ID=576

Credit for Data: from a vicious to a **virtuous** circle?



See The Value of Research Data: Metrics for datasets from a cultural and technical point of view http://www.knowledge-exchange.info/datametrics and Knowledge Exchange Workshop: 'Making Data Count': http://www.knowledge-exchange.info/Default.aspx?ID=576

Building a Culture of Data Citation

http://www.ands.org.au/guides/data_citation_poster.pdf

Developments in Data Citation

- **ICSU, *International Council for Science*, Statement on 'Open access to scientific data and literature and the assessment of research by metrics', Sept 2014 http://bit.ly/icsu-OA-statement**
- Endorses the OECD Principles and Guidelines on Access to Data from Publicly Funded Research (2007)
- Recommendation 4: '**Science publishers and chief editors of scientific publications should require authors to provide explicit references to the datasets underlying published papers, using unique persistent identifiers.** They also should require clear assurances that these datasets are deposited and available in trusted and sustainable digital repositories. **Citing datasets in reference lists using an accepted standard format should be considered the norm.**'

ICSU
International Council for Science

Need for Improved Metrics

- DORA, the San Francisco ***Declaration on Research Assessement*** http://dmm.biologists.org/content/early/2013/05/16/dmm.012955.full.pdf
- **Research assessment should include value and impact of all research outputs (including data and code) as well as qualitative indicators.**
- **Shift to article-based metrics, rather than journal-based metrics.**

- ICSU, International Council of Science Statement: http://bit.ly/icsu-OA-statement
- Endorses DORA **(Recommendation 11)**
- **Recommendation 10:** In research evaluation and assessment, metrics should be regarded as an aid, and not a substitute, for good decision-making. They should not normally be used in isolation to assess the performance of researchers, to determine appointments, or to distribute funds to individuals or research groups, for which expert review is indispensable.

'Do not use journal-based metrics, such as journal impact factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion or funding decisions.'

Developments: Journal Data Policies

- **Dryad Joint Data Archiving Policy, Feb 2010: http://datadryad.org/jdap**
- This journal **requires, as a condition for publication**, that data supporting the results in the paper should be archived in an appropriate public archive, such as GenBank, TreeBASE, Dryad, or the Knowledge Network for Biocomplexity.

- **PLOS Data Availability Policy, revised Feb 2014:** http://www.plosone.org/static/policies.action#sharing
- PLOS journals require authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exceptions.

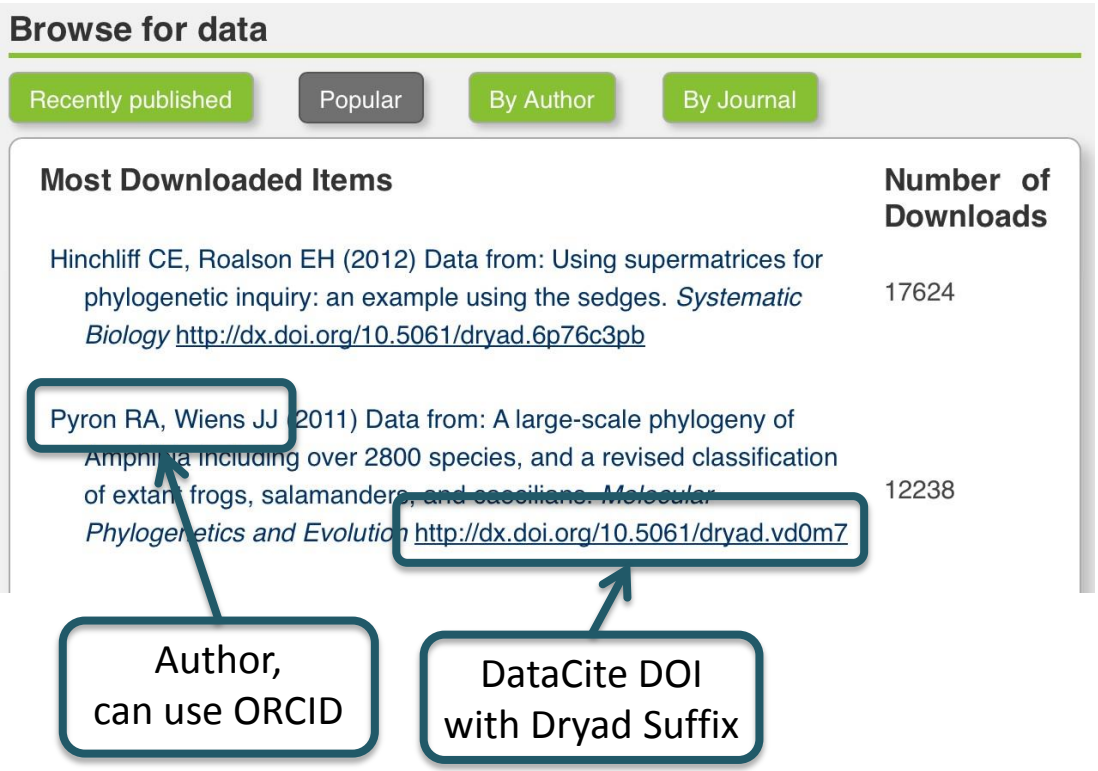
DRYAD

PLOS | ONE

How are we doing?

- 'Public Data Archiving in Ecology and Evolution: How Well Are We Doing?' DOI: 10.1371/journal.pbio.1002295
- 100 datasets in Dryad were surveyed: '56% were incomplete, and 64% were archived in a way that partially or entirely prevented reuse.'
- Of the 56, most had small amounts of data missing, suggested that involuntary.
- 'The most common pitfalls that affected data reusability were inadequate metadata, the use of proprietary and non-machine-readable file formats (e.g., data tables archived as PDF and word documents and failure to archive raw data.'
- 'Ecologists and evolutionary biologists receive little or no training in data management and may be unfamiliar with the best practices for proper data archiving': struggled to share effectively and in accordance with good practice.
- **Small, simple improvements can dramatically increase the reusability of archived data with minimal time or monetary investments.**
- Training and incentives.

Data Citations, DOIs and ORCIDs



ODIN

- Assists with associating data citations unambiguously with a given author.
- Increases visibility of data sources.
- **ODIN ORCID/DataCite Claim tool allowing authors to link dataset records with DataCite DOIs to their ORCID Profile.**
- Adoption of PIDs and interoperability of those PIDs is an important enabler of data citation, data sharing.

ORCID Auto-Update

Researchers can now set their ORCID profiles to automatically update with any published article or dataset associated with their ORCID.

http://blog.datacite.org/auto-update-has-arrived/

ICSU CODATA

FORCE11 Data Citation Implementation Group (DCIG)

- Clark, Starr et al, 'Achieving human and machine accessibility of cited data in scholarly publications', *PeerJ Computer Science* 1:e1 https://dx.doi.org/10.7717/peerj-cs.1
- Responsibilities of data archives in making the components for a citation available and ensuring persistence of source, landing page.
- Responsibilities of journals in ensuring data is credited through citation and including data citations and included in the reference list and so visible to citation indexes.

What should publishers do?

- **JATS 'Journal Article Tag Set', XML tags for journal articles.** This has recently been modified to better facilitate citing data and including data citation in the reference list: Citing Data in Journal Articles using JATS https://www.force11.org/sites/default/files/d7/project/882/citing-data-in-jats-2015-06.pdf
- Main issues for publishers are (thanks Tim Clark):
  - Adopt JATS 1.1d3 or a later revision, to support data citation based on the JATS model;
  - Ensure that new JATS elements translate properly into publisher HTML and PDF representations;
  - Adopt a standard list of appropriate repositories for varying kinds of data;
  - Require authors to provide a valid accession number from an approved repository, as a condition for completing peer review;

Thank you for your attention!

Slide credits: Louise Corti

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