CASRAI

More research, less paperwork, better information

Webinar Agenda

- Welcome & Introductions - 5 min
- CASRAI **Overview** - 10 min
- Toward a **Shared International Research Data Glossary** - 15 min
- Toward agreements on **Dataset Level Metrics** - 15 min
- Questions - 15 min

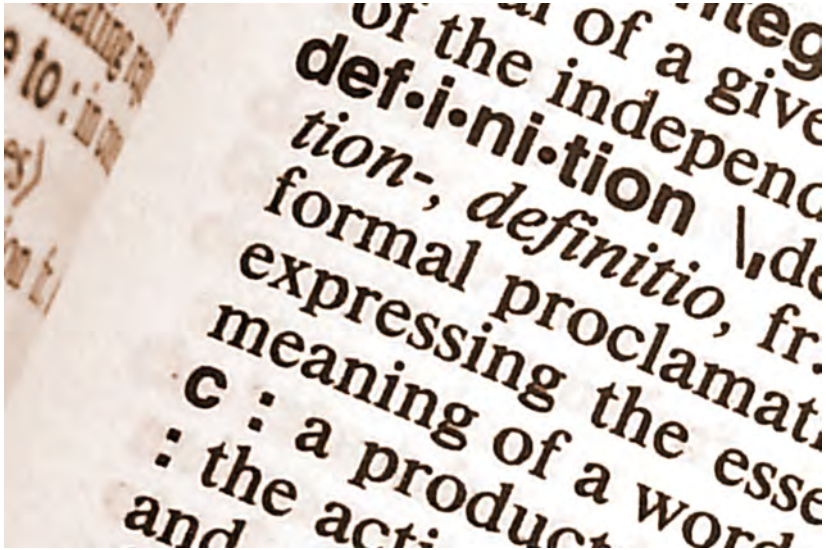
Objectives:

- Greater understanding of CASRAI
- Building communities-of-interest

1. CASRAI Overview

What is CASRAI?

- International nonprofit membership initiative **led by research institutions** and their partners
- Adapting principles and practices of **open standards and data governance** to develop/maintain ‘standard information agreements’
- To collectively **reduce paperwork** and **improve information quality**



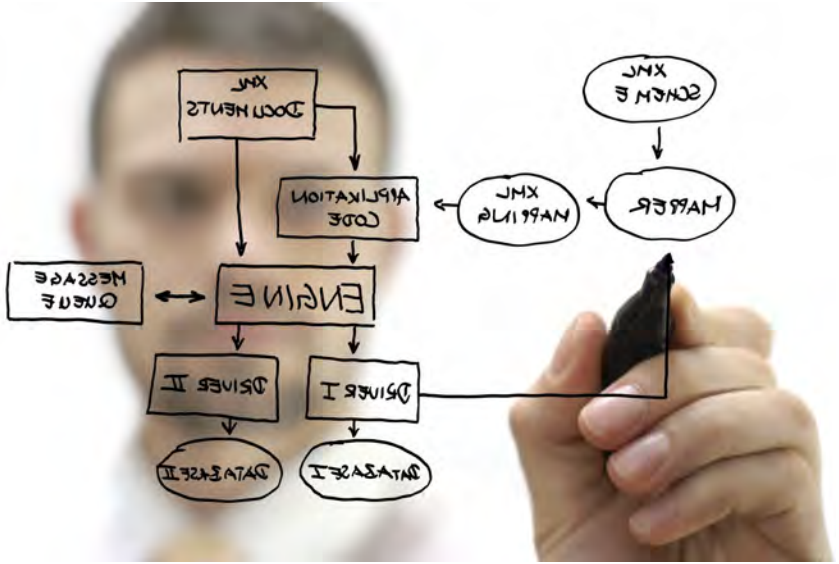
Six Contributing Factors to the Problem

1. Duplication
2. Complexity
3. Change
4. Feasibility
5. Comparability
6. Maturity

All the above further **compounded** by the fact that researchers and institutions have **multiple** information ‘requesters’

What we do

- Convene key users (**suppliers and consumers**) of research information
- Develop ‘user’ agreements (**pre-technology**) on definitions (glossaries) and report formats
- Submit drafts for **open** community **review**/revisions
- Publish (with unique identifiers) to a **web-based dictionary**
- Expressed also in **canonical XML**



2. Shared International Glossary

Tower of Babel

- One universal language
- A big, tall tower
- Aspirations of great things
- Failure, dispersion
- Confusion, conflict
- Many languages
- Division

BUNGEE - “a part” (spoken by the Métis in Manitoba)

- Many languages

  Scottish, Gaelic, Orcadian, Cree, Ojibee

- Draw from each other

- Find a common understanding

A Research Data Glossary

IT IS NOT:

- An all encompassing super mega dictionary to replace any other glossary or dictionary.

IT IS:

- A practical tool to enable trans-disciplinary teams
- Facilitate communication - individuals and working groups
- A forum for discussion and development of new terms

Research Data Glossary

STABILITY:

- Create a stable and sustainably governed glossary of community accepted terms and definitions,

RELEVANCE:

- Keep it relevant by maintaining it as a ‘living document’ that is updated when necessary.

The Need

- Difficulties resulting from a lack of a shared vocabulary.
- People from diverse domains need to be involved in research data management:
  - Information technology, Library science, Computer infrastructure, Computer science, Software engineering, Computer security, Statistics, Data science (which is itself an emerging discipline), Administration, Management, Project management, Publishing.
  - All of the domain sciences that produce research data.
- Multiple specialty lexicons developed over time; same terms sometimes used in different ways.

Defining Success

- The most important characteristic of a glossary is not necessarily that it contains the best possible terms and definitions, but that
- The measure of success will be that people
  - agree with and use the terms & definitions; and,
  - agree to sustain a process of iteratively evolving the glossary.

Development Methods

- Glossaries collected from government, community, and social media online sources.
- Criteria for consideration: ease of availability, relevance, and anticipated popularity.
- Automated methods for term extraction were not used.
- The glossary is not meant to replace machine-useable sources of terms definitions (ie semantic web, RDF technologies) but, rather, to provide a ‘human-useable’ starting point for agreement.

Sources included in Pilot

- RDC’s 1st Edition of a glossary (published in May 2014);
- WhatIs.com (purchased by TechTarget in 1999)
- Research Data Alliance (RDA) glossary under development by the Foundation and Terminology working and interest groups;
- Advancing Open Standards for the Information Society (OASIS) glossary.

Pilot Reach

- Emailed to International and national stakeholders.
- LinkedIn, Twitter
  - Estimated reach was 145,648 accounts,
  - The target audience resulting from organizational retweets exceeded 124,906
  - The top Tweet was Mozilla Science
  - The top mention was Open Science
- Glossary landing page hits from 1280 unique IP addresses between August 25 and September 8, 2015.
  - Click-through rate to term definition pages was 52% (668 hits).

Get involved

We hope you will join us to advance work in 2016:

- http://casrai.org

BioCaddie Webinar
14 April 2016

CODATA and the CASRAI-RDC Research Data Glossary

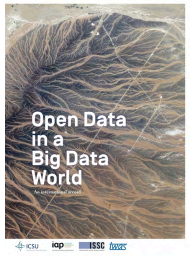
Dr Simon Hodson
Executive Director, CODATA
www.codata.org

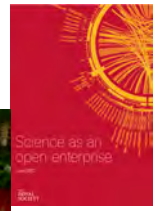
CODATA

Principles, Policies and Practice

DC¹ Data Citation Principles

ICSU International Council for Science

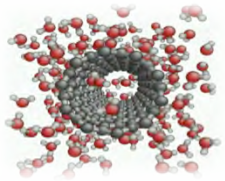




Capacity Building

The Abdus Salam International Centre for Theoretical Physics

Frontiers of Data Science



ubiquity press open scholarship

Data Science Journal

SciDataCon 2016, 11-13 Sept, Denver, CO.

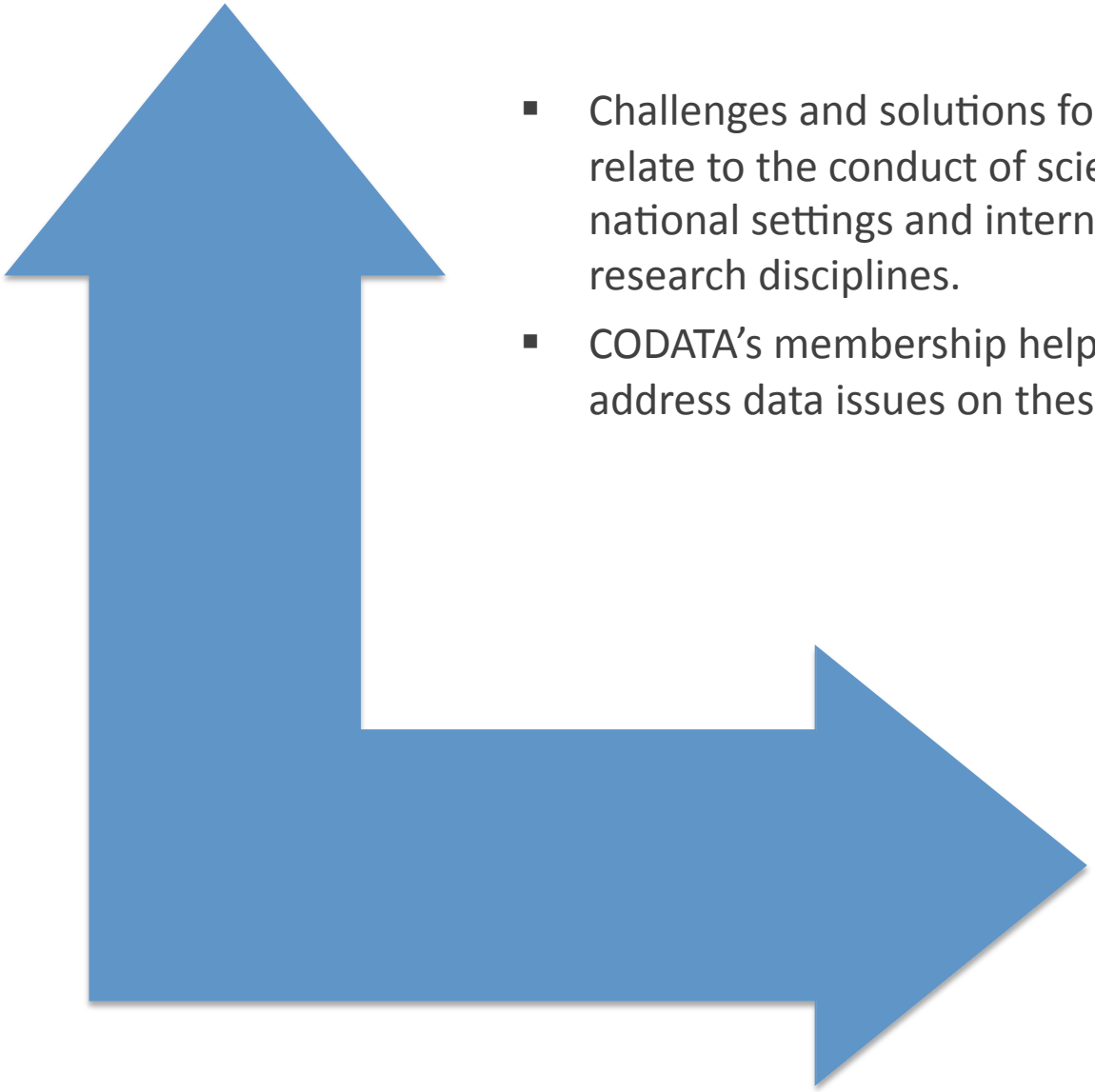
INTERNATIONAL DATA WEEK 2016
WWW.INTERNATIONALDATAWEEK.ORG

Research Data: challenges and stakeholders

National Research Systems

CODATA National Members

National Academies of Science or Data Organisations



- Challenges and solutions for data issues relate to the conduct of science in national settings and international research disciplines.
- CODATA's membership helps us to address data issues on these two axes.

Scientific Disciplines

CODATA International Scientific Union Members

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

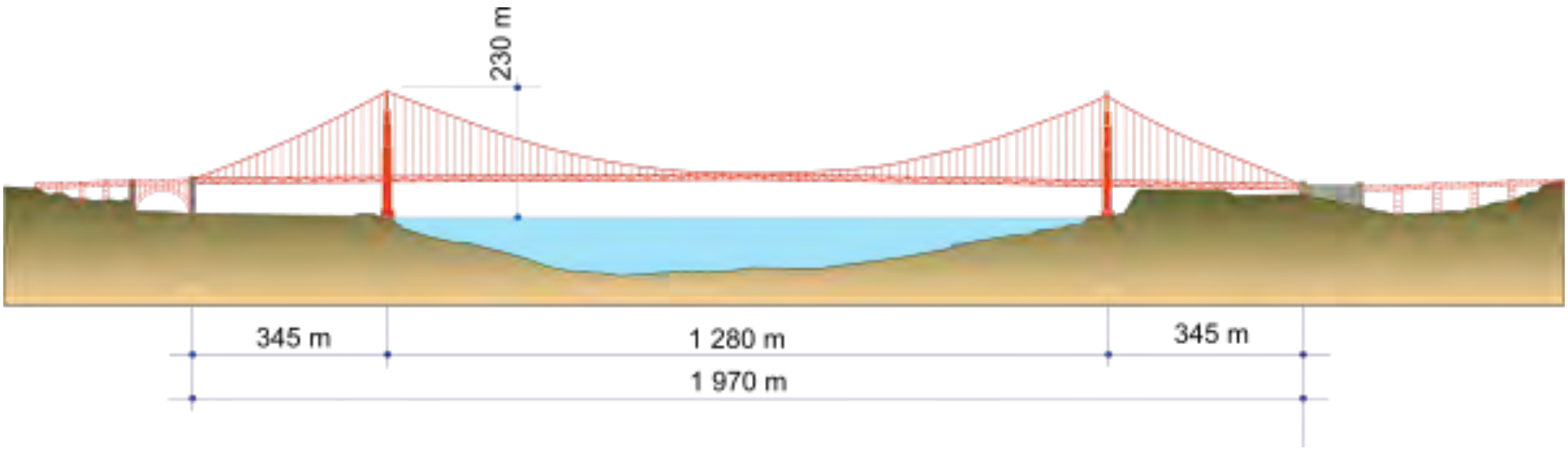
2014 CODATA RECOMMENDED VALUES OF THE FUNDAMENTAL CONSTANTS OF PHYSICS AND CHEMISTRY NIST SP 959 (Aug 2015)

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	$927.400\,9994(57) \times 10^{-26}$	$\mathrm{J\ T^{-1}}$

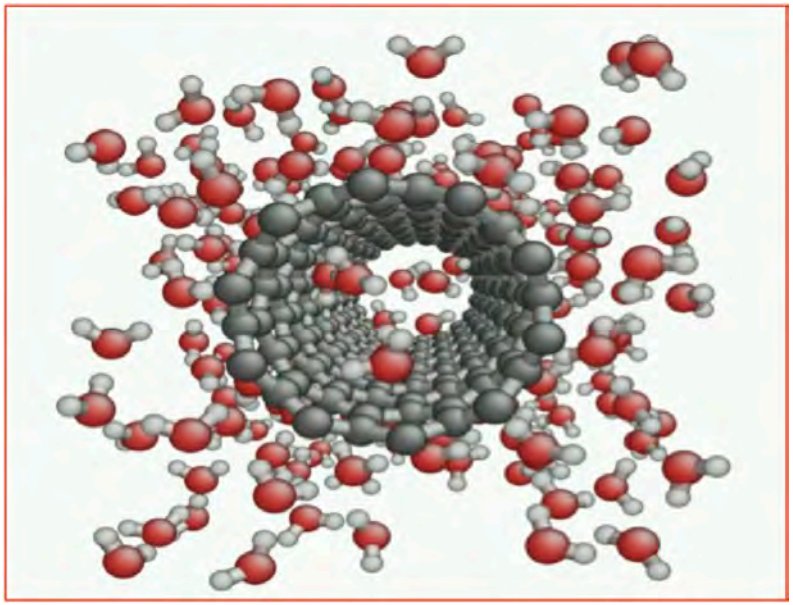
CODATA MO

- Leverage links with ICSU and ISUs; and with National Committees.
- Convenes researchers, data experts/scientists, standards bodies, international organisations.
- Task Groups or Working Groups develop frameworks, White Papers, authoritative and influential reports.
- Work with stakeholders to encourage uptake.



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



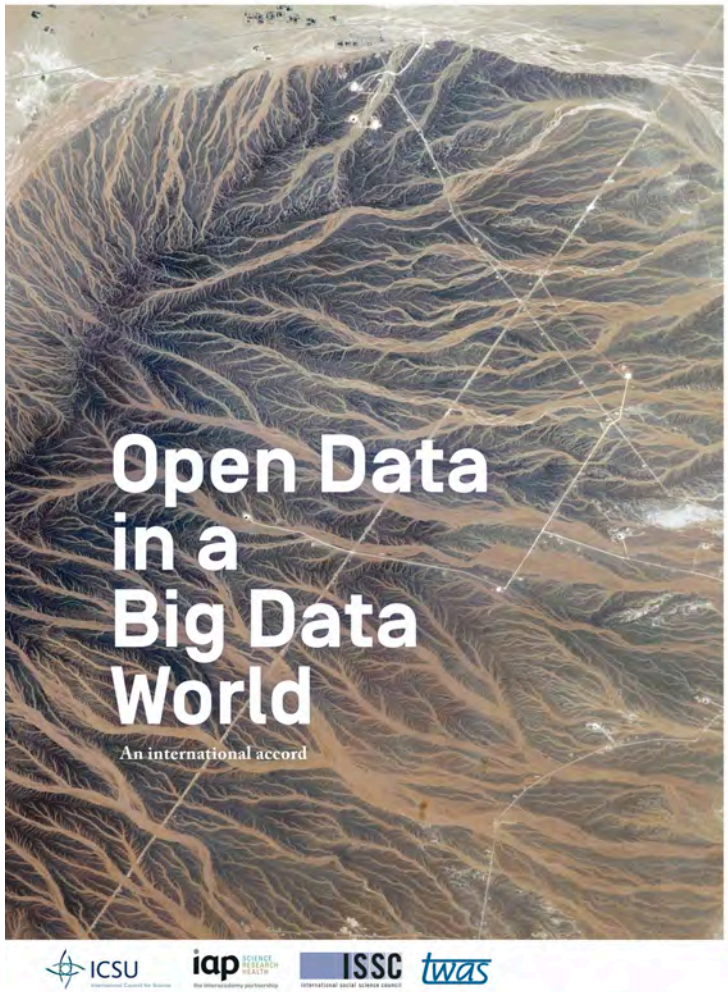
CODATA-ICSU Workshop:
http://www.codata.info/Nanomaterials/Index-agenda-Nanomaterial.html

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

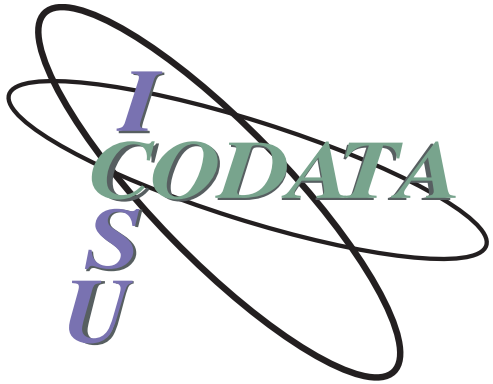
Uniform Description System v.01, Feb 2015:
http://dx.doi.org/10.5281/zenodo.20688

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

Why a glossary for research data terms?



- Science is international!
- Increasing need for international coordination on research data activity.
- Essential to have agreement over terms!
- CODATA has a long history and mission to provide means of international coordination on development of consensus, standards etc.
- Admire the mechanisms, method and process designed by CASRAI.
- CODATA can bring in an international community relating to the academies and unions.



ICSU

International Council for Science

Thank you for your attention!

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

3. Research Dataset Level Metrics

Rationale of the Community-of-Interest

Two main concerns:

- Researchers should get a **career benefit** from sharing data
  - How to demonstrate impact?
- Researchers can’t tell if their data are **reusable**
  - How to get quality assurance?

Can we make the solutions scalable with metrics?

Aims of the Community-of-Interest

Define sets of metrics that can be used:

- As a starting point for decisions about **impact** and **quality**
- As a **motivator** for doing the right thing

Potential Use Cases

- A **publisher** displaying metrics for data underlying an article
- A **repository** giving depositors feedback about how their data are being used
- A **funder** wanting to evaluate the data outputs of funded research
- A **university** wanting to highlight impacts of data outputs
- A **university** wanting to use data outputs in recruitment or promotion decisions

Related Work: Making Data Count

- CDL/PLOS/DataONE project, Oct 2015 - Oct 2015
  - Found researchers most interested in **citations** and **downloads**
- Repurposed Lagotto (PLOS Article Level Metrics) for data
  - Existing support for counting downloads, formal citations
  - Harvesting **informal citations** (links, IDs) from open access corpora
  - Monitoring **mentions** in social media, DataCite metadata, etc.
- Further reading
  - Lagotto on GitHub
  - ‘Making data count’ http://doi.org/10.1038/sdata.2015.39
  - ‘When counting is hard’ https://blog.datacite.org/when-counting-is-hard/

Related: RDA/WDS Publishing Data Bibliometrics WG

- Oct 2014 - March 2016
- Found all stakeholder groups most interested in **citations** and **downloads**
- Draft recommendations:
  - Funders to mandate data deposit and citation
  - Stakeholders to agree on how to use repository statistics
  - Stakeholders to use multiple metrics, not rely on just one
- Further reading:
  - https://rd-alliance.org/groups/rdawds-publishing-data-bibliometrics-wg.html

Related Work: NISO Altmetrics Group B

- Two-phase project, June 2015 - Jan 2015 - March 2016
- Group B: ‘non-traditional research outputs and identifiers’
- Draft recommendations:
  - Implement Force11 data citation principles
  - Use COUNTER standards to count **human**-initiated downloads
  - Define new standards for counting (or not) **machine** interactions
- Further reading:
  - http://niso.org/topics/tl/altmetrics_initiative/

Quality Dataset-Level Metrics for Repositories WG

- Use case:
  - A generalist data repository applies a set of metrics to a deposited dataset to determine whether it meets a given level of quality
- Aim:
  - Define a clear set of metrics such repositories could use
- Activity:
  - Two telecons
  - Workshop at BioMed Bridges closing symposium
  - Relaunch in 2016 for a three-month sprint

Quality Considerations

- Does the submission have integrity?
- Was it collected according to an accepted methodology?
- Is the raw-to-result processing reproducible?
- What kinds of format(s) does it use?
- Is it properly structured and labelled?
- Does it include added discipline-associated metadata?
- Has it been documented sufficiently for reuse?
- Has the dataset been appropriately licensed?
- Have appropriate access restrictions been applied?
- Has it already passed an expert review?
- Does it compare favourably with ‘known good’ datasets?

Get involved

We hope you will join us to advance work in 2016:

- http://casrai.org

Questions/Discussion