Research Data Management: Trends and International Context

Kathleen Shearer

Consultant, Research Data Canada

Research Associate, Canadian Association of Research Libraries

Executive Director, Confederation of Open Access Repositories

m.kathleen.shearer@gmail.com

@kathleeshearer

Open Science

- Sharing and access to all types of research outputs including data (when appropriate)
- Transparency and validation of research findings
- Open peer review & open usage metrics
- Equitable flow of knowledge

➢ Began with open access to publications, but moving to data and other types of research outputs

➢ Parallels to Open Government/Open Data movement

What is driving this trend?

1. Verification, reproducibility and transparency of scientific results
2. Better, more efficient research
3. New scientific discoveries through re-use datasets
4. Greater diffusion of knowledge and therefore social and economic impacts through application of research outputs
5. And because we can...

Figure ES.1 Average data storage cost for consumers

1998-2012, per Gbit
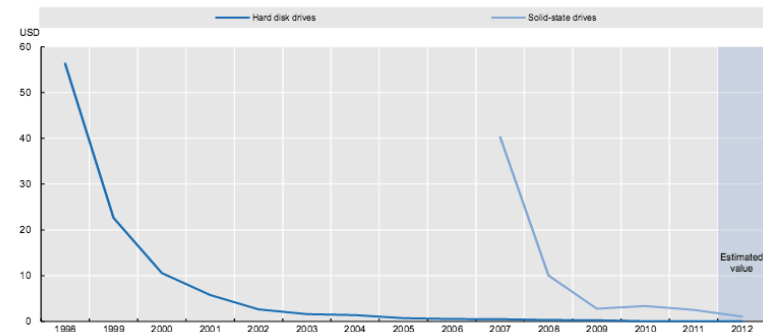
Source: OECD (2014), *Measuring the Digital Economy: A New Perspective*, OECD Publishing, Paris.

The Poster Child:

- Sequencing centres agreed to make results available the same day

- (1990-2013) Finished coding the human genome 2 years ahead of time

- Contributes to treatment to diseases

- Billions of dollars in economic spin-off

- Transformed the way biomedical research is conducted

# Policies and laws are important levers to move open science forward

LEVER

FULCRUM

Man lifting a stone
with a lever
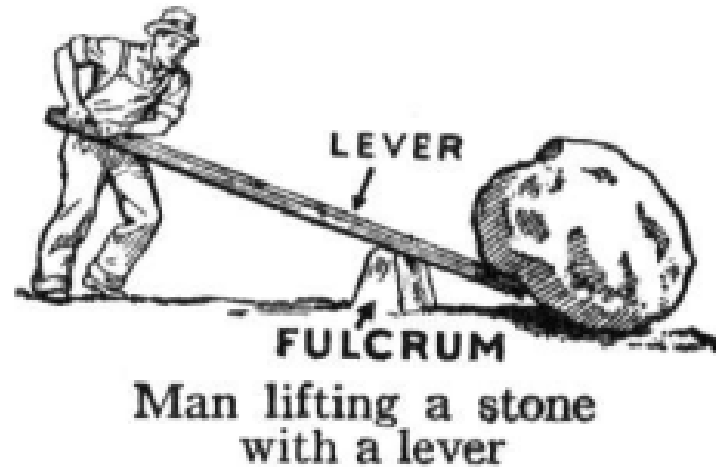
Simple Machines: Lever

Moving towards openness

Open science: a hot issue for OECD and non-OECD countries

Number of countries reporting that the situation has recently substantially changed in the policy area, compared with other STI policy areas or instruments
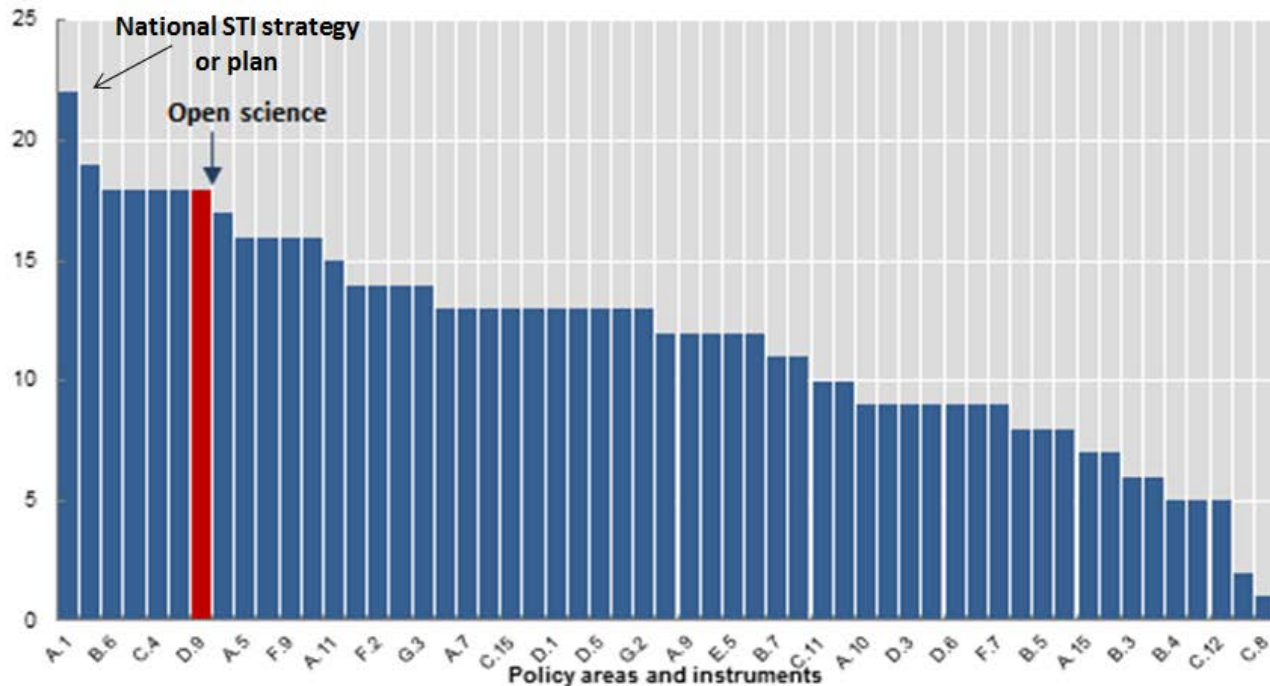
Note: Simple counts do not account for the magnitude and impact of policy changes.

Source: Country responses to the STI Outlook policy questionnaire 2014.

G8 Science Ministers, June 2013

UK, Russia, Germany, Japan, Italy, Canada, France, US and EC

- **Expanding Access to Scientific Research Results**

We endorse the principle that increasing access to the peer-reviewed, published results of publicly funded published research will accelerate research, drive innovation, and benefit the economy.

- **Open Scientific Research Data**

We are committed to openness in scientific research data to speed up the progress of scientific discovery, create innovation, ensure that the results of scientific research are as widely available as practical, enable transparency in science and engage the public in the scientific process.

Canada's Action Plan on Open Government 2014-2016
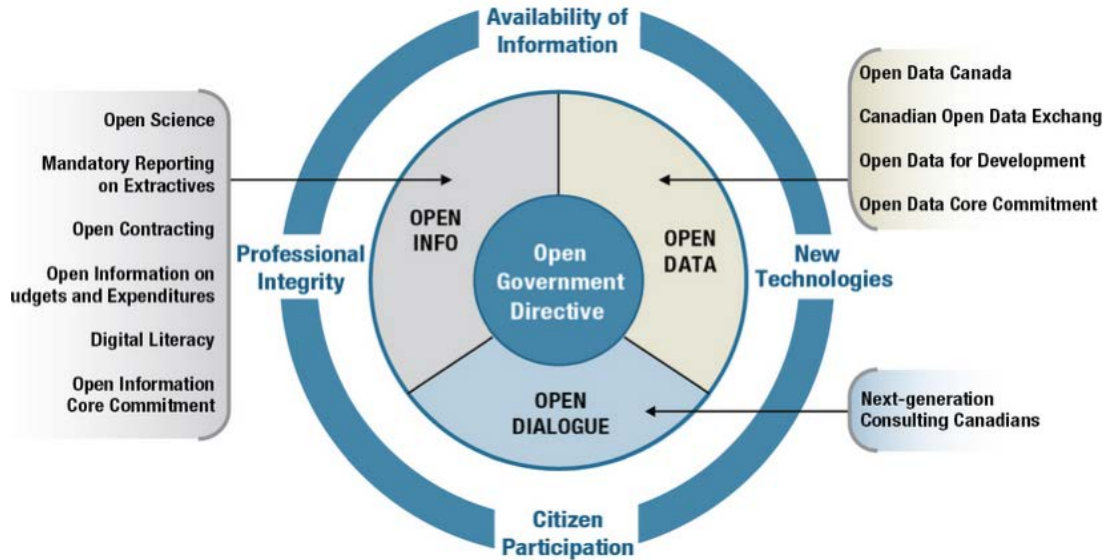
"...the Government of Canada will establish a government-wide approach to Open Science to increase access to federally-funded scientific publications and data."

International Policy Environment

World Funders by Data Archiving Policy Type

- Required (42 = 27%)
- Encouraged (18 = 11%)
- (No Policy) (93 = 59%)

JULIET 27-Oct-2015

Total = 157 funders

From Sherpa-Juliet: http://www.sherpa.ac.uk/juliet/stats.php?la=en&mode=simple

Typical Elements of a Data Policy

Policy requirements	
Data quality and standards	Investigators are required to adhere to international standards to enable access and reuse in the discipline. Data documentation and metadata must accompany data so that the data is understandable by others.
Data access and sharing	Investigators are required to make data available to be shared (usually upon publication of results or shortly thereafter, although some agencies do allow embargo periods). Requirements for deposit of metadata into a local or national catalogue
Data retention and preservation	Data should be retained for a certain time limit, where possible, investigators must deposit their data in a long-term archive to ensure the preservation of their data.
Data management plans	Research proposals must include a Data Management Plan in proposal.

Exceptions

Common exceptions to policies	
Privacy	The rights and privacy of individuals who participate in research must be protected at all times. Thus, data intended for broader use should be free of identifiers that would permit linkages to individual research participants and variables that could lead to deductive disclosure of the identity of individual subjects.
Traditional knowledge	Where local and traditional knowledge is concerned, rights of the knowledge holders shall not be compromised.
Data of a sensitive nature	Where data release may cause harm, specific aspects of the data may need to be kept protected (for example, locations of nests of endangered birds or locations of sacred sites).
Intellectual property/Data ownership	It may be necessary on occasion to delay publication for a short period to allow time for applications to be drafted.

Data Management Plans

portage
Shared stewardship of research data

Sections	Questions
Data Collection	- What types of data will you collect, create, link to, acquire and/or record? - What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data? - What conventions and procedures will you use to structure, name and version-control your files to help you and others better understand how your data are organized?
Documentation and Metadata	- What documentation will be needed for the data to be read and interpreted correctly in the future? - How will you make sure that documentation is created or captured consistently throughout your project? - If you are using a metadata standard and/or tools to document and describe your data, please list here.
Storage and Backup	- What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it? - How and where will your data be stored and backed up during your research project? - How will the research team and other collaborators access, modify, and contribute data throughout the project?
Preservation	- Where will you deposit your data for long-term preservation and access at the end of your research project? - Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.
Sharing and Reuse	- What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final). - Have you considered what type of end-user license to include with your data? - What steps will be taken to help the research community know that your data exists?
Responsibilities and Resources	- Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible. - How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator? - What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?
Ethics and Legal Compliance	- If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project? - If applicable, what strategies will you undertake to address secondary uses of sensitive data? - How will you manage legal, ethical, and intellectual property issues?

But we need more than just policies
Research Data Management is like a three-legged stool...

Culture

Infrastructure
and services

Policies

Data sharing practices

From Wiley's Research Data Insights Survey, 2014
http://exchanges.wiley.com/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-don't/
2,250 responses from around the world

Data tension

Human tension and/or stress related to the sharing or release of data resulting from concerns about: (a) unknowns about users, uses, and what users will learn from the data before the data producers themselves learn it; (b) what users will learn from the data; (c) data quality; (d) data traceability (or lack thereof); (e) potential requests for additional documentation and metadata; (f) potential questions concerning methodology used to produce the data; (g) lack of resources to support data sharing; (h) governance; (i) social or political interests and impact; (j) data ownership; (k) the desire to ëhold backí data to give researchers the time to publish articles based on those data; and/or (l) perceived risk of data misuse or misinterpretation.

Discussion

BIG DATA!

90% of world's data has been generated over last two years

http://www.sciencedaily.com/releases/
2013/05/130522085217.htm
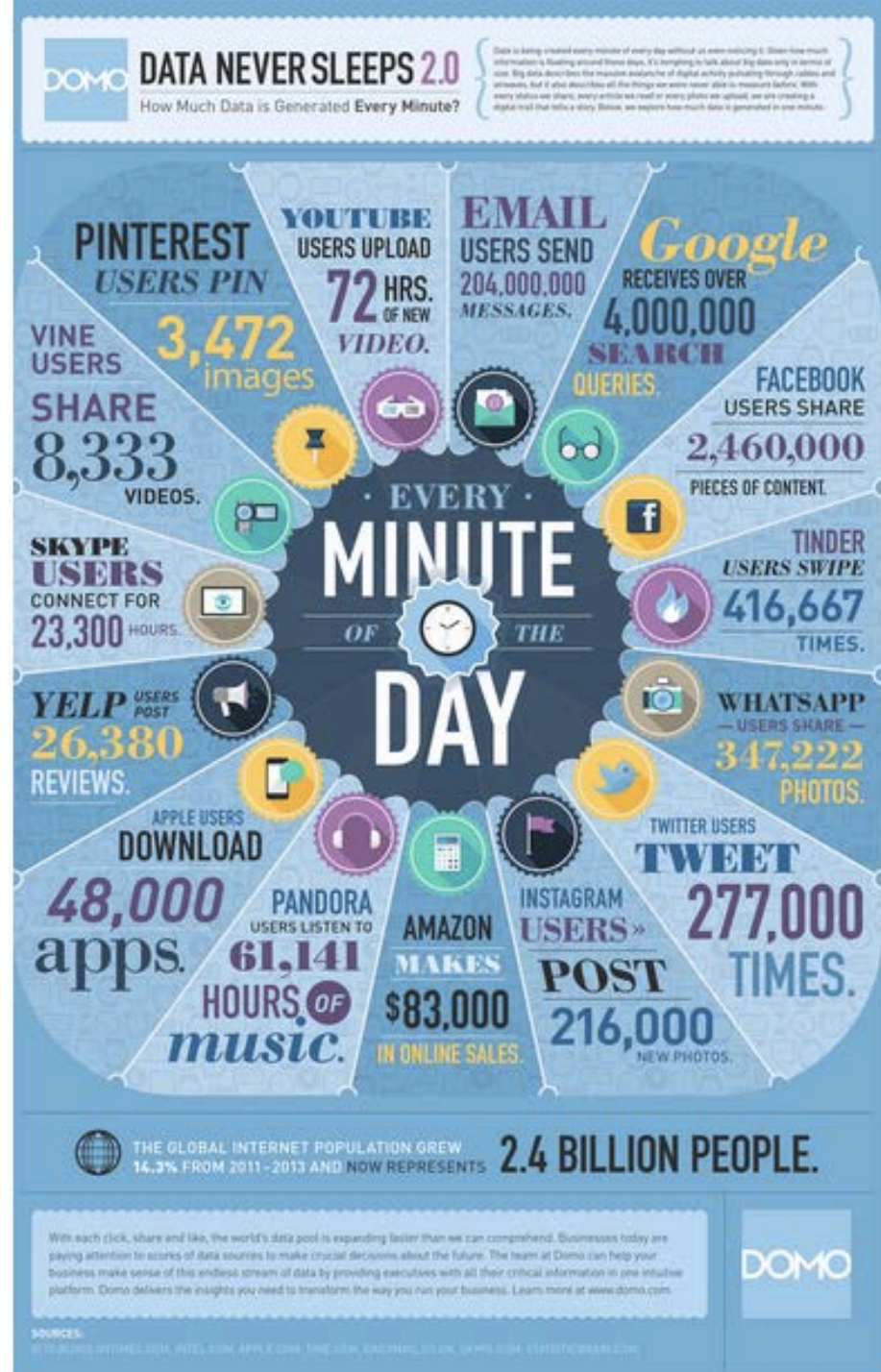
Image: Erik Fitzpatrick licensed CC BY 2.0

Infrastructure and Services
Data sharing requires good management across the data lifecycle
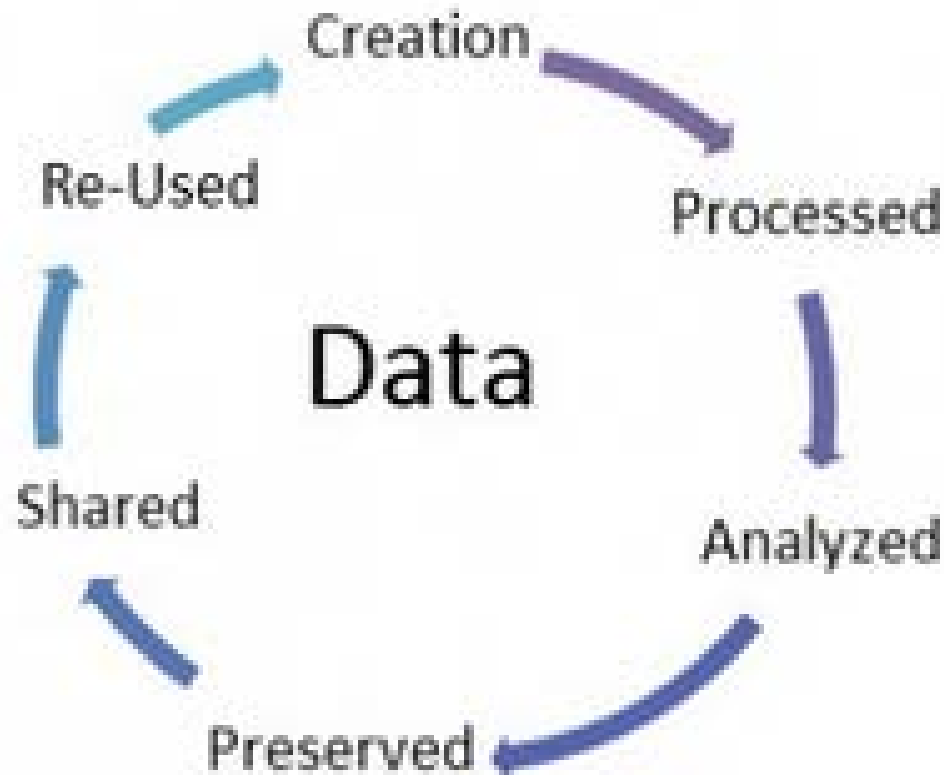
The data landscape

These services still only support a small portion of the research datasets produced by researchers around the world!

The data landscape

The 2011 survey by *Science*, found that 48.3% of respondents were working with datasets that were less than 1GB in size and over half of those polled store their data only in their laboratories. *Science* 11 February 2011: Vol. 331 no. 6018 pp. 692-693 *DOI:* 10.1126/science.331.6018.692

What is the size of the largest data set that you have used or generated in your research?

1 TB

100 GB

1 GB

7.6% >1 TB
12.1% 100 GB–1 TB
32.0% 1–100 GB
48.3% <1 GB
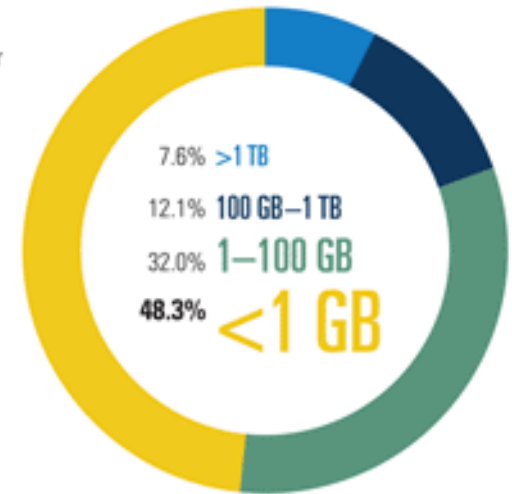
Where do you archive most of the data generated in your lab or for your research?

" Even within a single institution **there are no standards for storing data**, so each lab, or often each fellow, uses ad hoc approaches. "

0.5% It is not stored
50.2% Our Lab
38.5% University servers
7.6% Community repository
3.2% Other
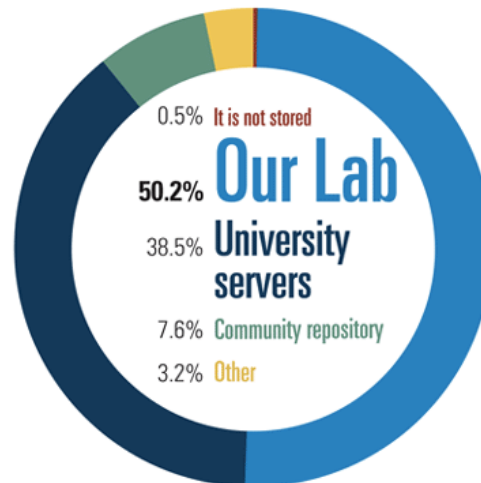
Institutional Role: UK

In this section

Roadmaps to EPSRC expectations on research data

In April 2011, the Engineering & Physical Sciences Research Council (EPSRC) set out its research data management expectations for institutions in receipt of EPSRC grant funding, which included the development of an institutional 'Roadmap'.

Related Information

Five DCC blogs posts on developing roadmaps

- A Journey Shared
- Turning Roadmaps to Action
- Navigating the Potholes
- The Essential Data Roadmap
- Developing a Roadmap for RDM

Higher Education Institutions and their Roadmaps:

- Oxford Brookes University published their roadmap in May 2012.
- Keele University made available their draft roadmap on 1 June 2012.
- University of Bath announced the publication of its roadmap on 7 June 2012. The Glossary & Definitions section in particular has been welcomed by

Emergence of National Data Services