Wagging the Long Tail of Research Data

Kathleen Shearer, Executive Director, COAR

Co-chair, RDA Long Tail for Research Data Interest Group

Co-chair RDA Libraries and Research Data (soon to be IG)

Project Coordinator, Project ARC, a library based initiative to develop a national network for RDM in Canada

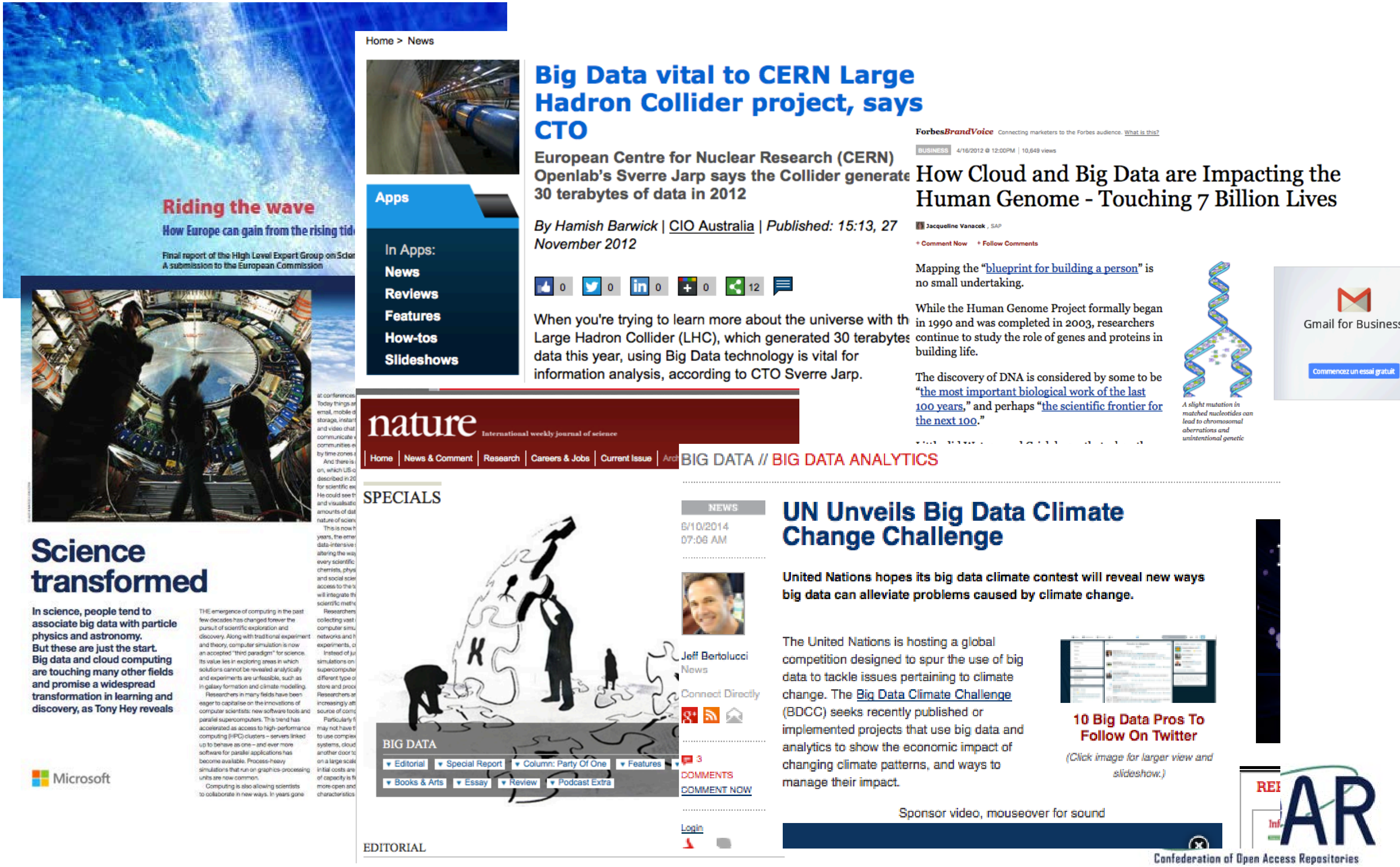
Confederation of Open Access Repositories

About The Confederation of Open Access Repositories (COAR)

- Over 100 institutional members from around the world on five continents
- Mission: to create a global network of open access repositories in support of research
- Community of practice & an international voice for the OA repository community
- Major issue is **interoperability** (repository-repository AND repository-other systems)
- To date, mainly focused on institutional role in managing and providing open access to publications
- These services are evolving/expanding to include the management of **research data**

COAR
Confederation of Open Access Repositories

“Big data” is all the rage



Riding the wave

How Europe can gain from the rising tid

Final report of the High Level Expert Group on Scien
A submission to the European Commission

Science transformed

In science, people tend to associate big data with particle physics and astronomy. But these are just the start. Big data and cloud computing are touching many other fields and promise a widespread transformation in learning and discovery, as Tony Hey reveals

THE emergence of computing in the past few decades has changed forever the pursuit of scientific exploration and discovery. Along with traditional experiment and theory, computer simulation is now an accepted “third paradigm” for science. Its value lies in exploring areas in which solutions cannot be revealed analytically and experiments are unfeasible, such as in galaxy formation and climate modelling.

Researchers in many fields have been eager to capitalise on the innovations of computer scientists: new software tools and parallel supercomputers. This trend has accelerated as access to high-performance computing (HPC) clusters – servers linked up to behave as one – and ever more software for parallel applications has become available. Process-heavy simulations that run on graphics-processing units are now common.

Computing is also allowing scientists to collaborate in new ways. In years gone

Microsoft

Home > News

Big Data vital to CERN Large Hadron Collider project, says CTO

European Centre for Nuclear Research (CERN) Openlab's Sverre Jarp says the Collider generate 30 terabytes of data in 2012

By Hamish Barwick | CIO Australia | Published: 15:13, 27 November 2012

When you're trying to learn more about the universe with th Large Hadron Collider (LHC), which generated 30 terabytes data this year, using Big Data technology is vital for information analysis, according to CTO Sverre Jarp.

ForbesBrandVoice Connecting marketers to the Forbes audience. What is this?

BUSINESS 4/16/2012 @ 12:00PM | 10,649 views

How Cloud and Big Data are Impacting the Human Genome - Touching 7 Billion Lives

Jacqueline Vanacek , SAP

+ Comment Now + Follow Comments

Mapping the “blueprint for building a person” is no small undertaking.

While the Human Genome Project formally began in 1990 and was completed in 2003, researchers continue to study the role of genes and proteins in building life.

The discovery of DNA is considered by some to be “the most important biological work of the last 100 years,” and perhaps “the scientific frontier for the next 100.”

A slight mutation in matched nucleotides can lead to chromosomal aberrations and unintentional genetic

SPECIALS

EDITORIAL

BIG DATA // BIG DATA ANALYTICS

UN Unveils Big Data Climate Change Challenge

United Nations hopes its big data climate contest will reveal new ways big data can alleviate problems caused by climate change.

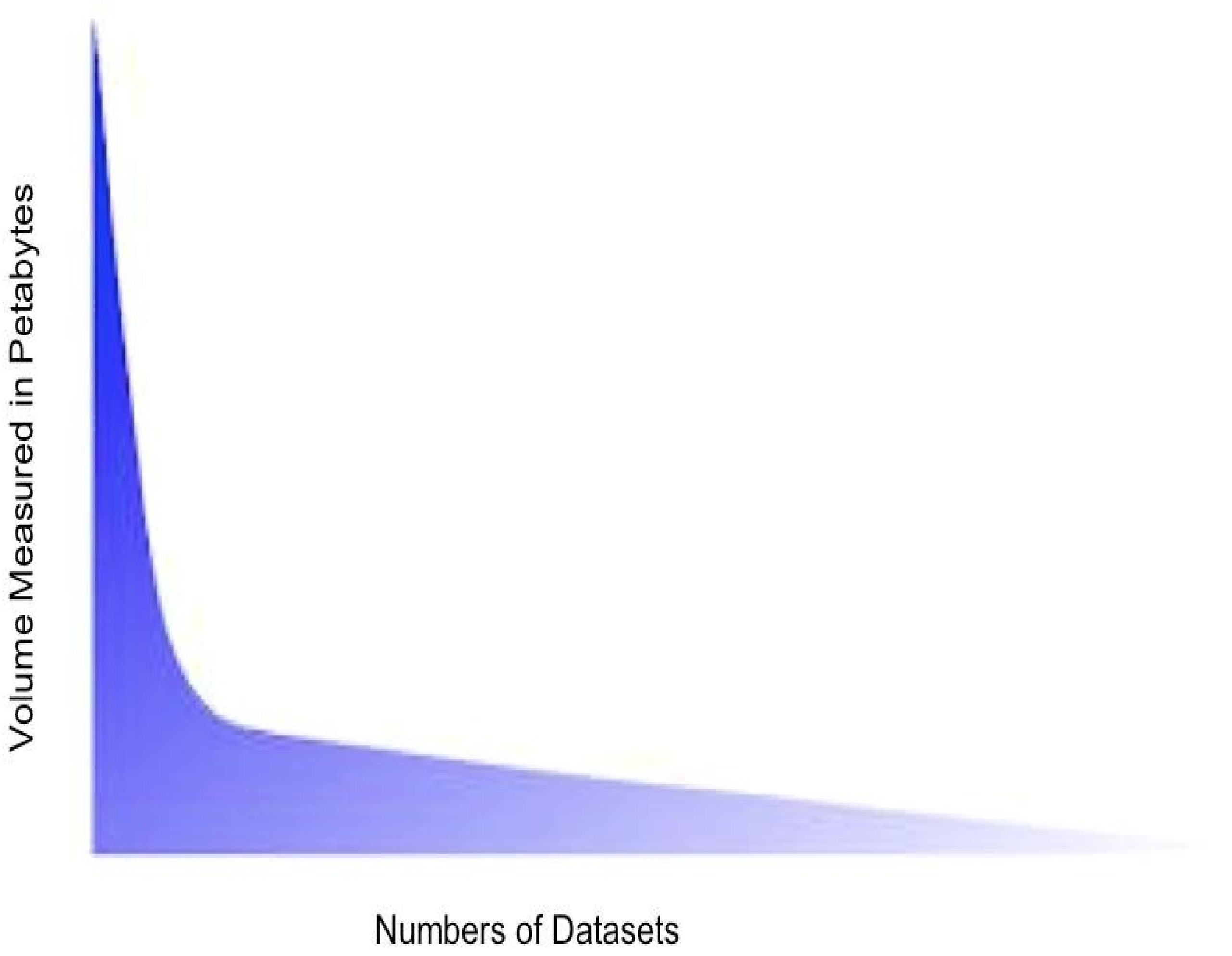
The United Nations is hosting a global competition designed to spur the use of big data to tackle issues pertaining to climate change. The Big Data Climate Challenge (BDCC) seeks recently published or implemented projects that use big data and analytics to show the economic impact of changing climate patterns, and ways to manage their impact.

10 Big Data Pros To Follow On Twitter

(Click image for larger view and slideshow.)

Sponsor video, mouseover for sound

But, the majority of datasets produced through research are part of the "Long Tail of Research Data"



(Image from Chuck Humphrey: OpenAIRE-COAR Conference, Athens 2014)

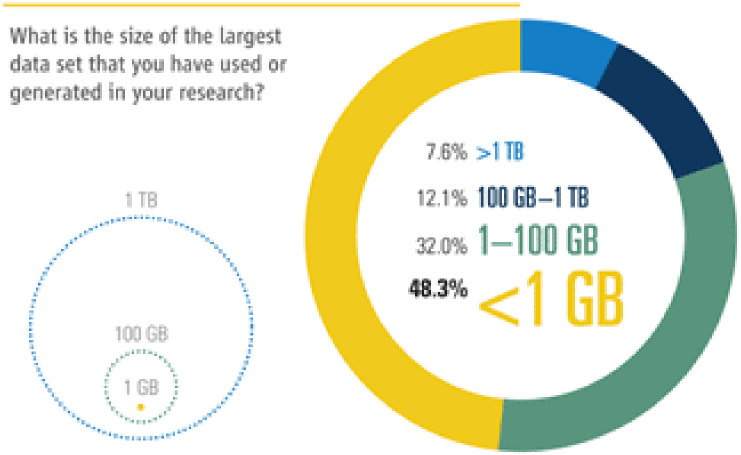
Characteristics of Long Tail Research Data

Head	Tail
Homogeneous	Heterogeneous
Large	Small
Common standards	Unique standards
Integrated	Not-integrated
Central curation	Individual curation
Disciplinary repositories	Institutional, general or no repositories

Adapted from: *Shedding Light on the Dark Data in the Long Tail of Science* by P. Bryan Heidorn. 2008

Long Tail of Research Data: small (…sometimes)

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



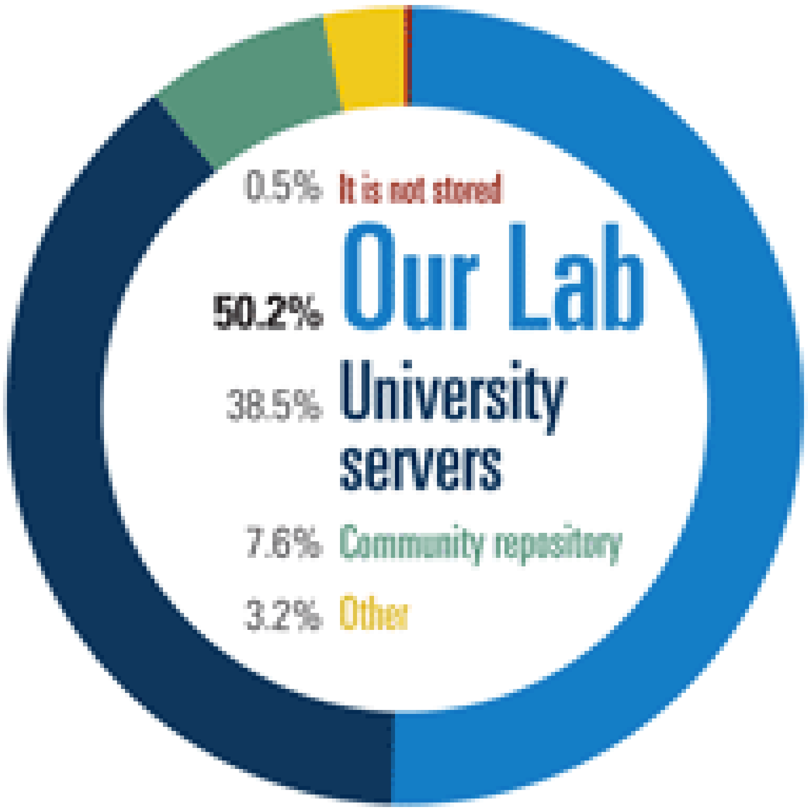
Long Tail of Research Data: heterogeneous

- A review undertaken by Cornell University of over 200 data “packages” (files related to arXiv papers) deposited into the Cornell Data Conservancy with there were 42 different file extensions for 1837 files across six disciplines. http://blogs.cornell.edu/dsps/2013/06/14/arxiv-data-conservancy-pilot/
- The Dryad Repository, which is a curated, general-purpose repository that collects and provides access to data underlying scientific publications reports a huge diversity of formats including excel, CVS, images, video, audio, html, xml, as well as “many uncommon and annoying formats”. The average size of the data package which they collect is ~50 MB. http://wiki.datadryad.org/wg/dryad/images/b/b7/2013MayVision.pdf
- According to the European Commission (EC) document, *Research Data e-Infrastructures: Framework for Action in H2020,* “diversity is likely to remain a dominant feature of research data – diversity of formats, types, vocabularies, and computational requirements – but also of the people and communities that generate and use the data.” http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/framework-for-action-in-h2020_en.pdf

Long Tail of Research Data: Institutional, general, domain or (often) no repositories

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.”



Science 11 February 2011: Vol. 331 no. 6018 pp. 692-693 *DOI:* 10.1126/science.331.6018.692

Confederation of Open Access Repositories

Long Tail of Research Data: some of the challenges

Data quality

- Determining quality and value of datasets
- Standards, metadata and norms differ significantly across disciplines

Discoverability

- diverse datasets are less discoverable because they are not found in a “go to” domain repository

Incentives

-why should researchers for deposit their data?

Business case

- why should organizations invest in the management of this data?

COAR
Confederation of Open Access Repositories

Long Tail of Research Data Interest Group

- Accepted as an RDA Interest Group in Summer 2013
- Over 90 members from around the world

Objectives

- To better understand the long tail
- To address some of the challenges involved in managing diverse datasets
- To share current practices, and develop best practices, for managing diverse data
- To work towards greater interoperability across repositories

Long Tail of Research Data Interest Group

Activities-to-date

- Survey of discovery metadata
- Discussion of strategies for improving discoverability of datasets

(All information is available on the interest group’s website)

Future activities

- evidence to incentivize researchers to deposit
- creating environments to make it easier for researchers to deposit their data,
- sharing practices about discovery,
- interoperability across repositories
- preservation planning

COAR

Confederation of Open Access Repositories

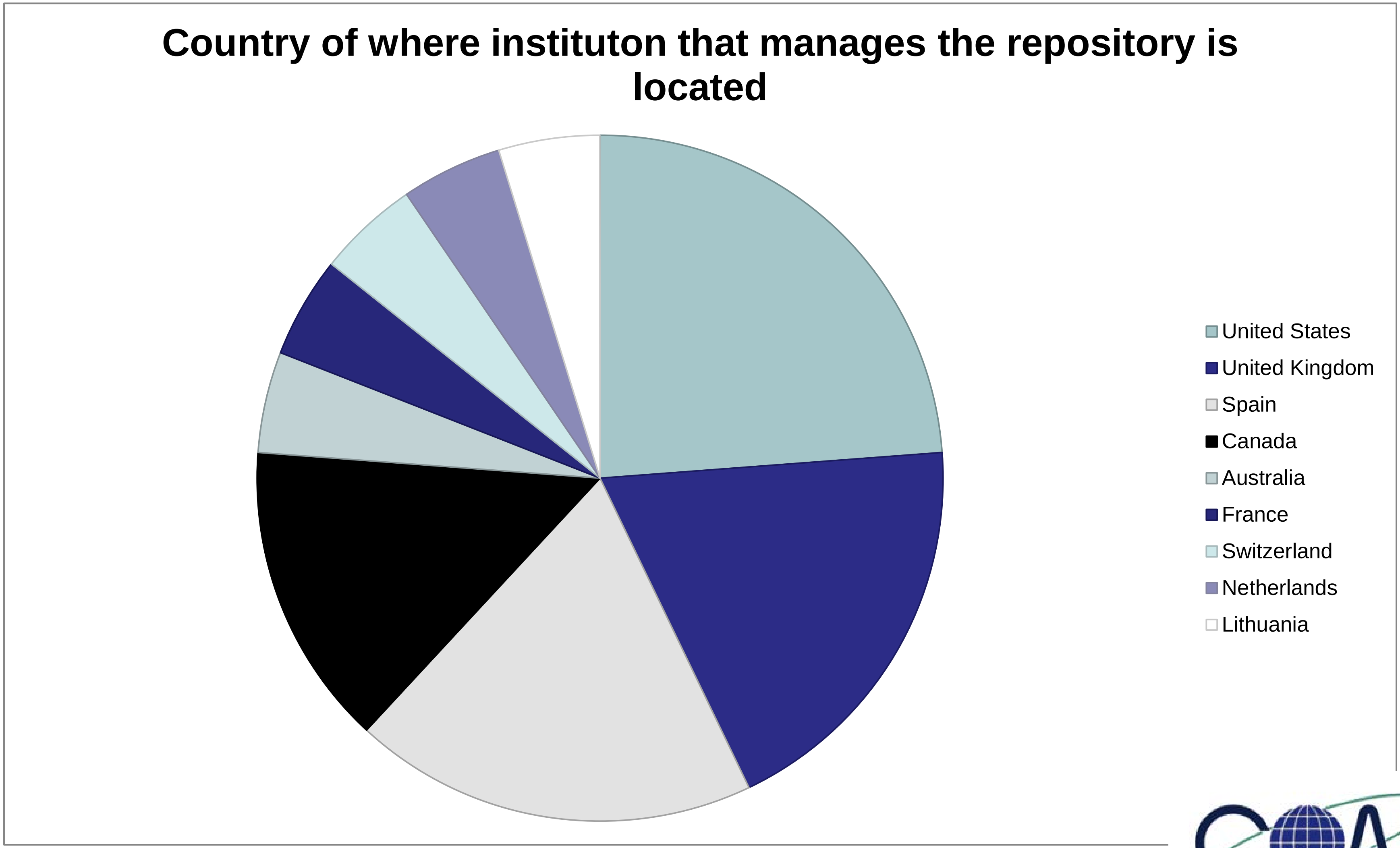
Survey of Current Practices for Discovery of Research Data

COAR
Confederation of Open Access Repositories

Survey of Current Practices for Discovery Metadata

- Purpose: to better understand the current practices in terms of discovery metadata
- Respondents: any repository collecting long tail data
- Undertaken from February 15 to March 7, 2014
- Recruited respondents via RDA mailing list and other research data list serves
- Over 60 responses, but only 30 full responses
- OBVIOUSLY not a representative sample, but an indication of which way the wind is blowing

Location of repository



What are the descriptive metadata standards used?

Repositories using a single schema

Dublin Core (9)
DataCite (3)
DDI Study-level metadata cf supra.
ISO19115 (Geographic Information Metadata)
MARC21
MODS metadata
RIF-CS

Repositories using more than one schema

DataCite and Dublin Core (3)
Dublin Core, Darwin Core, Prism
Dublin Core, EDM, ESE, QDC
Dublin Core, MARC21
dc, dcterms, geo/wgs84, FOAF, own extension ontology
MODS & DataCite Metadata Schema
Organic.Edunet IEEE LOM

In your opinion, is the metadata used in the repository sufficient to ensure discoverability of the datasets?

88% said yes, but…

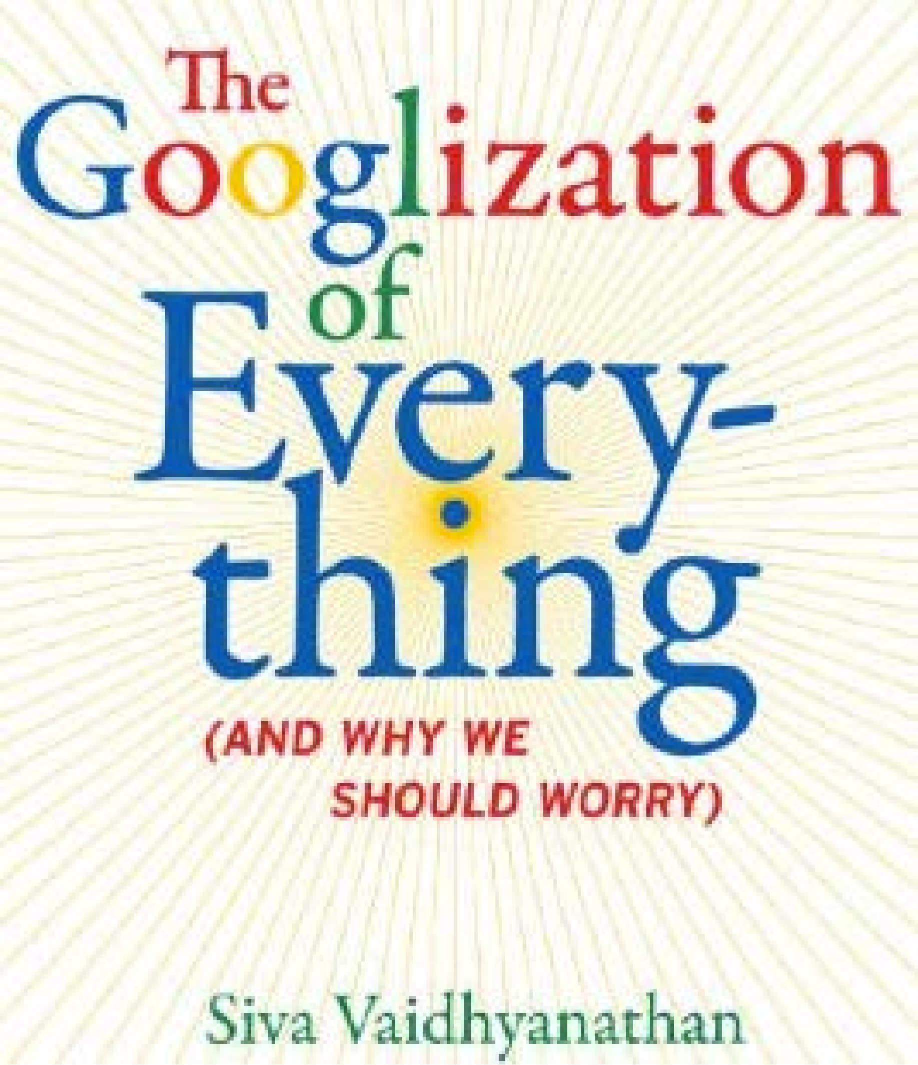
- Broadly speaking, and at a very high level, yes. If someone is looking for the data that supports a specific study, it is likely they will find it. However, if someone is looking for data with specific collection characteristics or other particularities then the metadata requires further enhancement.
- We aim to index metadata to aid discovery only. Metadata required to explore / reuse data will be stored with the data as a (non-indexed) object or stored in a separate, searchable database which links to the individual data objects in the repository (which may be at a sub-collection level). Data will also be found as the DOI will be included in publications related to the dataset.

In your opinion, is the metadata used in the repository sufficient to ensure discoverability of the datasets?

88% said yes, but…

- Data are discoverable within the repository because of limited repository scale, but once harvested and made available to search alongside tens of thousands of other datasets, the metadata are insufficient
- Precision is low because natural language metadata queries tend to entrain marginally relevant data sets due to weak associations in project descriptions and other broad fields.
- Fine for basic discoverability - richer discipline metadata would be nice but probably not feasible at this point

But we know, most most people use Google as their discovery tool



The Googlization of Every-thing

(AND WHY WE SHOULD WORRY)

Siva Vaidhyanathan

Strategies for improving the discoverability of datasets

- Linking data to publications
- Data citation- DOIs
- Build discovery layer that further describes data (landing pages)
- Attach or link to Data Management Plans (DMPs)
- Enable machine readability
- Data sets registries
- Data repository registries

Confederation of Open Access Repositories

Some concluding comments

- There is a growing interest in the management of long tail research data and institutions are recognizing they have a responsibility to manage research data
- Institutions can offer a sustainable, long-term solutions
- We already have a lot of expertise with metadata, preservation, and collaboration
- But, we need to work closely with data creators who have the disciplinary knowledge
- We have a lot to learn from the disciplinary communities about managing data
- We should heed the lessons learned from academic publishing (i.e. be wary of artificial measures of quality and impact)

COAR
Confederation of Open Access Repositories

kathleen.shearer@coar-repositories.org

COAR
Confederation of Open Access Repositories