

# Wagging the Long Tail

## Current Metadata Practices for Long Tail Research Data

Kathleen Shearer, Executive Director, COAR

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

Co-chair, RDA Libraries for Research Data Interest Group



# “Big data” is all the rage!



## 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 implications 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

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 generated 30 terabytes of data in 2012

By Hamish Barwick | [CIO Australia](#) | Published: 15:13, 27 November 2012

Facebook 0 Twitter 0 LinkedIn 0 + 0 RSS 12

When you're trying to learn more about the universe with the Large Hadron Collider (LHC), which generated 30 terabytes of 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 @ 12:20PM | 10,648 views

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

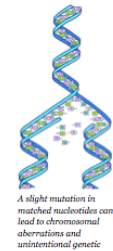
By Jacqueline Vanacek, SAP

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



**nature** International weekly journal of science

Home News & Comment Research Careers & Jobs Current Issue Arch

**BIG DATA // BIG DATA ANALYTICS**

NEWS 6/10/2014 07:06 AM

Jeff Bertolucci News

Connect Directly

**BIG DATA**

Editorial Special Report Column: Party Of One Features

Books & Arts Essay Review Podcast Extra

3 COMMENTS COMMENT NOW

Log in

## 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.)

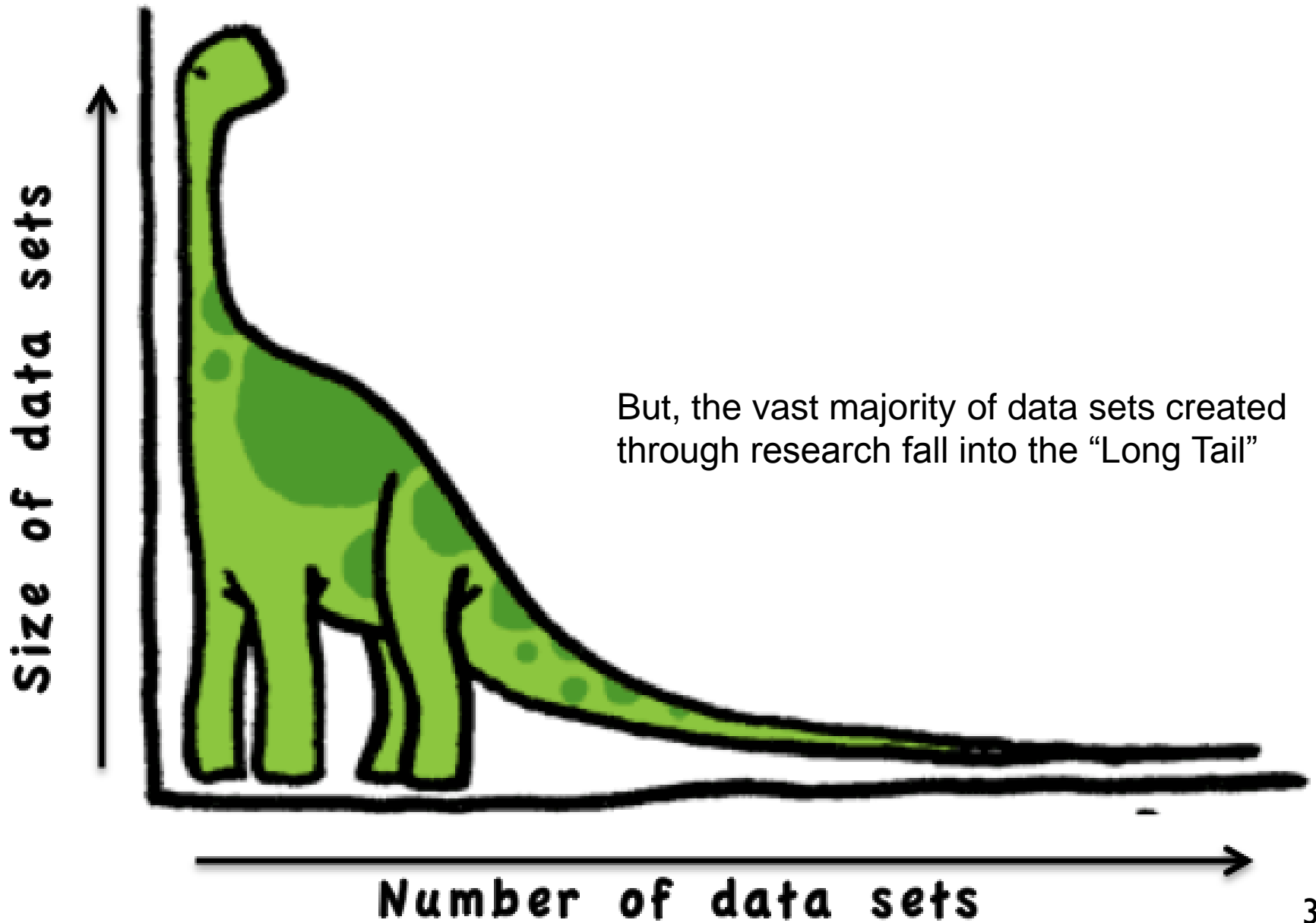
EUDAT - September 25, 2014 - Shearer

EDITORIAL



Commencez un essai gratuit





# The Long Tail

Head	Tail
Homogeneous	Heterogeneous
Interoperable, integrated	Non interoperable
Large	Small
Common standards	Unique standards
Central curation	Individual curation
Disciplinary repositories	Institutional, discipline, or most often, no repositories

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



# The Long Tail

- 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](http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/framework-for-action-in-h2020_en.pdf)



# The role of metadata

*Metadata remains the glue that holds information systems together. The better you manage your metadata, the better you serve your users. (Information Management, 2013)*

*Metadata quality is a vital factor for electronic interoperability. (Rousidis, et al. 2014)*

*Good quality, accurate and current metadata renders the research data more useful and accessible over the longer term. (Australian National Data Service)*



# In the context of Long Tail data, metadata is *critical* for discovery



[www.jolyon.co.uk](http://www.jolyon.co.uk)



# Survey of Discovery Metadata in Research Data in Repositories (Shearer, 2014)

Thanks to the following people for their input:

- Chuck Humphrey
- Stephen Marks
- Najla Rettberg
- Jochen Schirrwaggen
- Birgit Schmidt





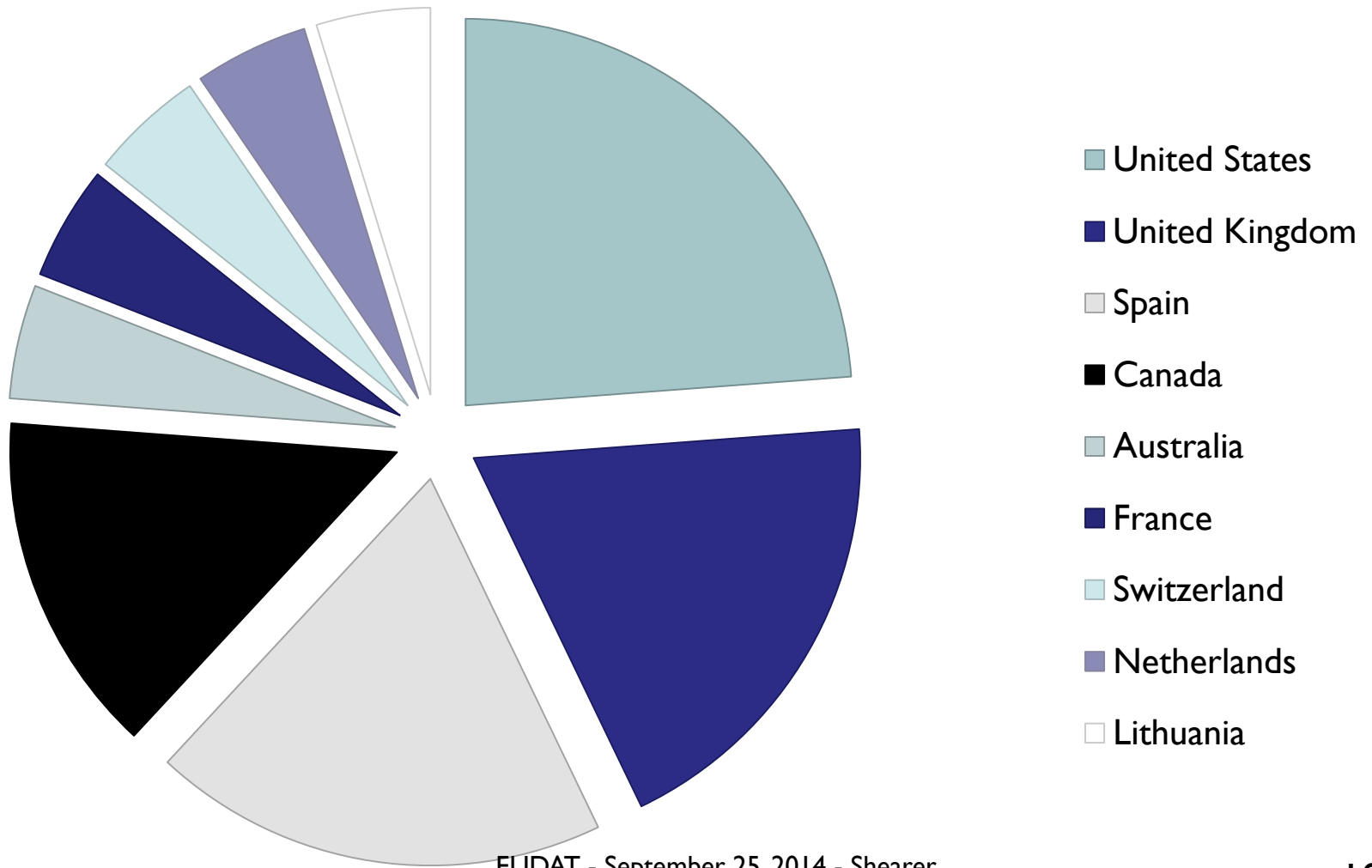
# Survey of Current Practices for Discovery Metadata

- 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

Country of where instituton that manages the repository is located



EUDAT - September 25, 2014 - Shearer

# Repository Platforms

## What repository platform are you using?

DSpace (9)

Fedora (3)

EPrints (2)

Islandora (2)

Locally developed (2)

Dataverse

Fez software (open  
source)

Greenstone Digital Library

Invenio

Metacat

Omeka

Postgres

RedBox and Mint

VIVO



## 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

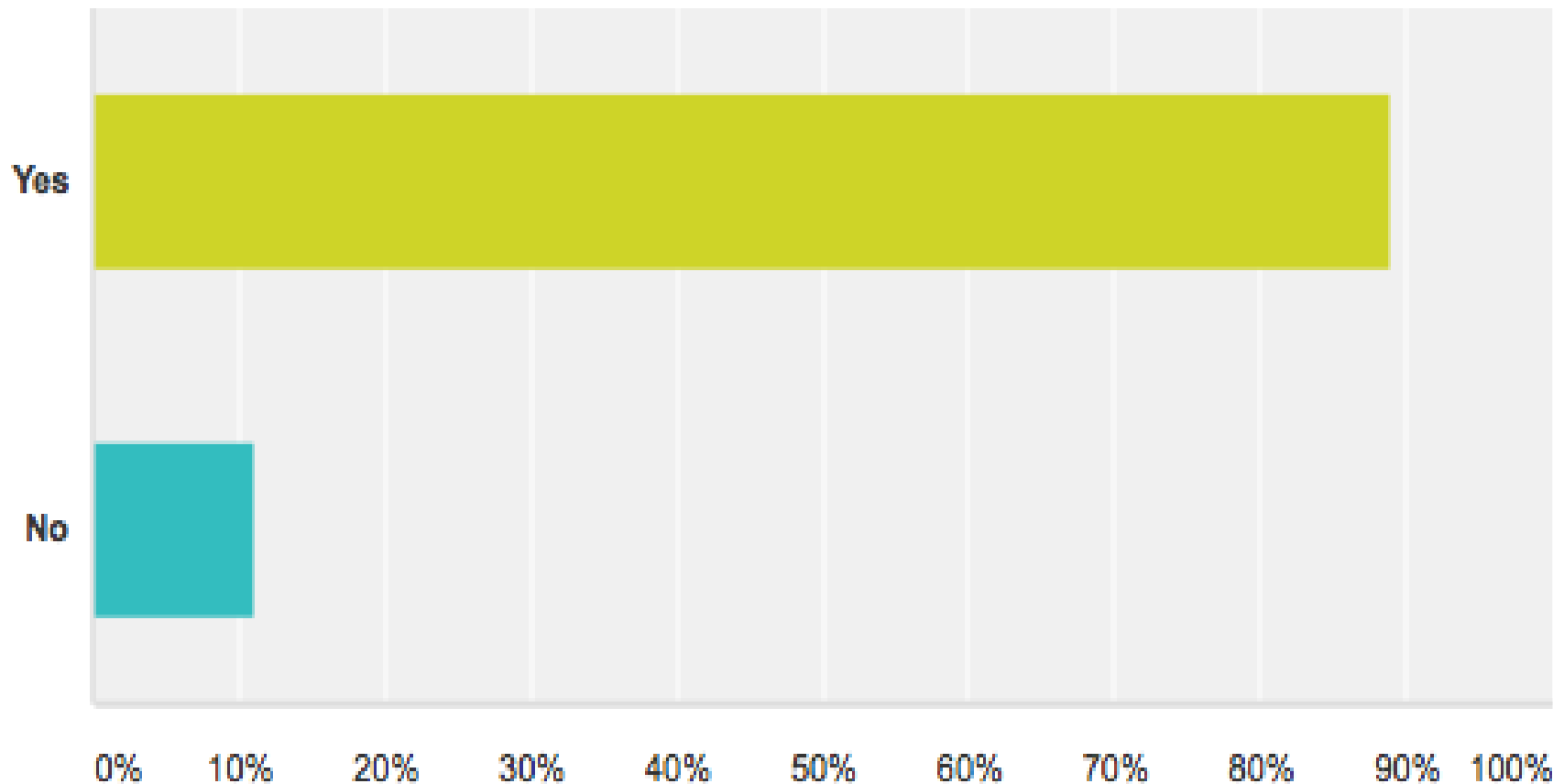


### Repositories using more than one schema

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?



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

## Yes...

- The few datasets that we have appear to be easily discoverable in Google.
- Compliance with OAI-PMH ensures discoverability; also integration into library search schemes.
- By requiring that data to be associated with related publications, this enriches the metadata.
- 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.



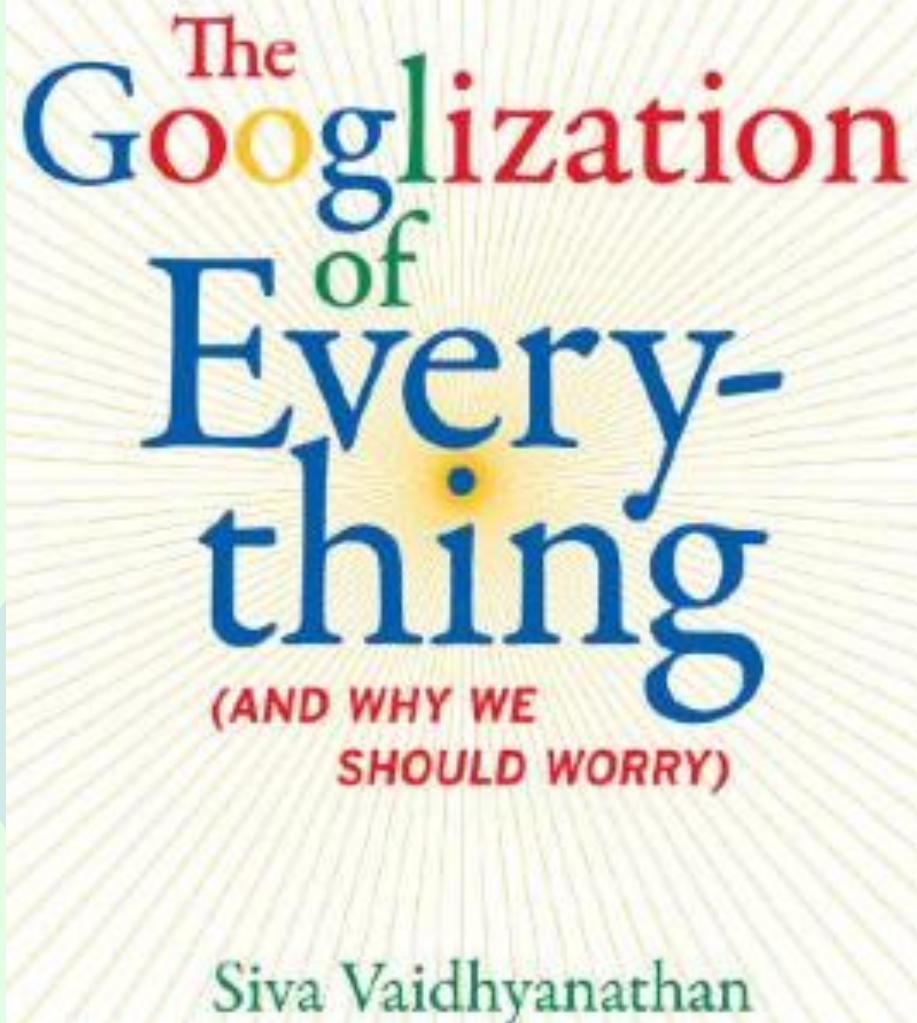
# In your opinion, is the metadata used in the repository sufficient to ensure discoverability of the datasets?

## 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.
- 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



Our conclusion: current metadata practices are sufficient for local discovery, however not for discovery through federated or external search services.



Yet, we know that most people use external services, such as Google as their main discovery tools.



# Some concluding comments

- Include DOIs and link to publications. This helps enrich existing metadata
- Ensure “long tail” repository platforms can incorporate different disciplinary schemas
- Splash or landing pages that describe datasets are very valuable for discovery - maybe data management plans could (eventually) be used for this?
- Working with researchers at the time of data production helps to improve the quality of metadata
- It will probably always be difficult to find the balance between lowering the barrier for deposit and capturing rich metadata



# Thanks. Questions?

Kathleen Shearer  
Executive Director, COAR

[kathleen.shearer@coar-repositories.org](mailto:kathleen.shearer@coar-repositories.org)

[www.coar-repositories.org](http://www.coar-repositories.org)

