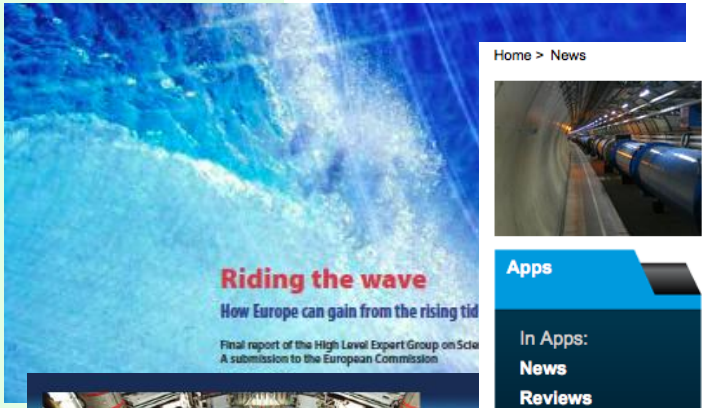
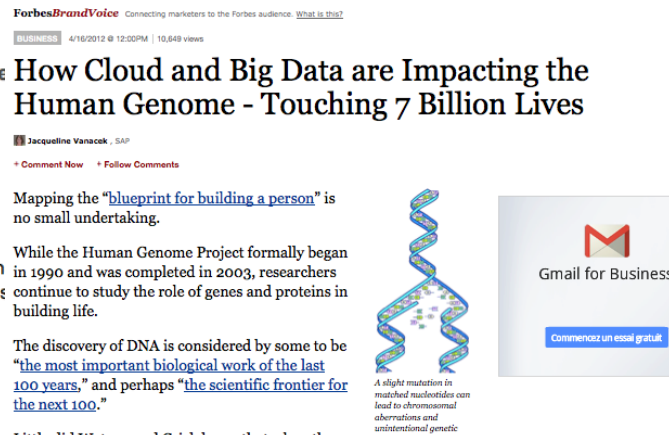
Wagging the Long Tail

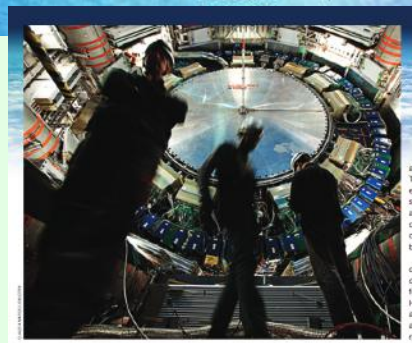
Current Metadata Practices for Long Tail Research Data

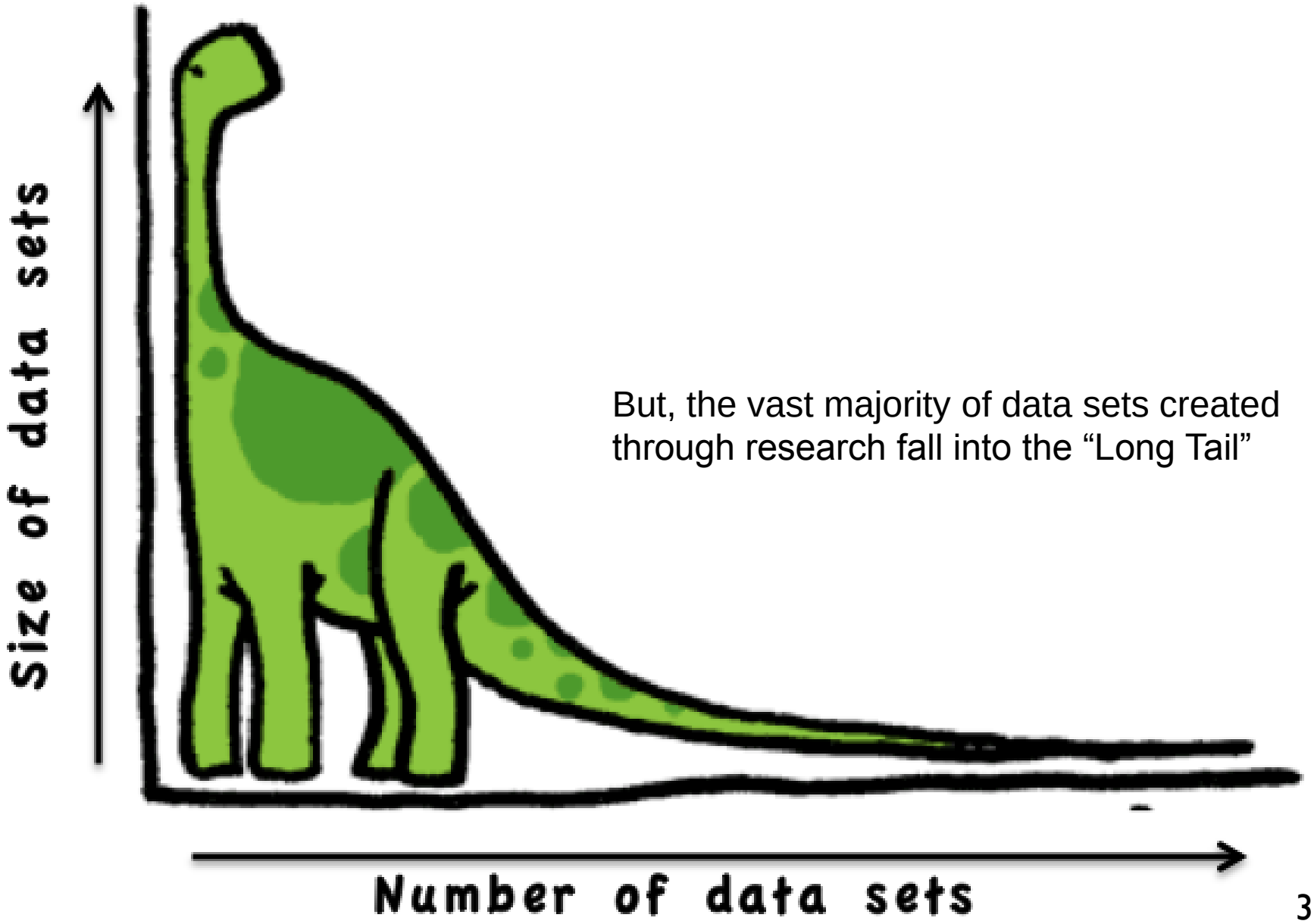
Kathleen Shearer, Executive Director, COAR
Co-chair, RDA Long Tail for Research Data Interest Group
Co-chair, RDA Libraries for Research Data Interest Group

“Big data” is all the rage!









But, the vast majority of data sets created through research fall into the “Long Tail”

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

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

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

Thanks to the following people for their input:

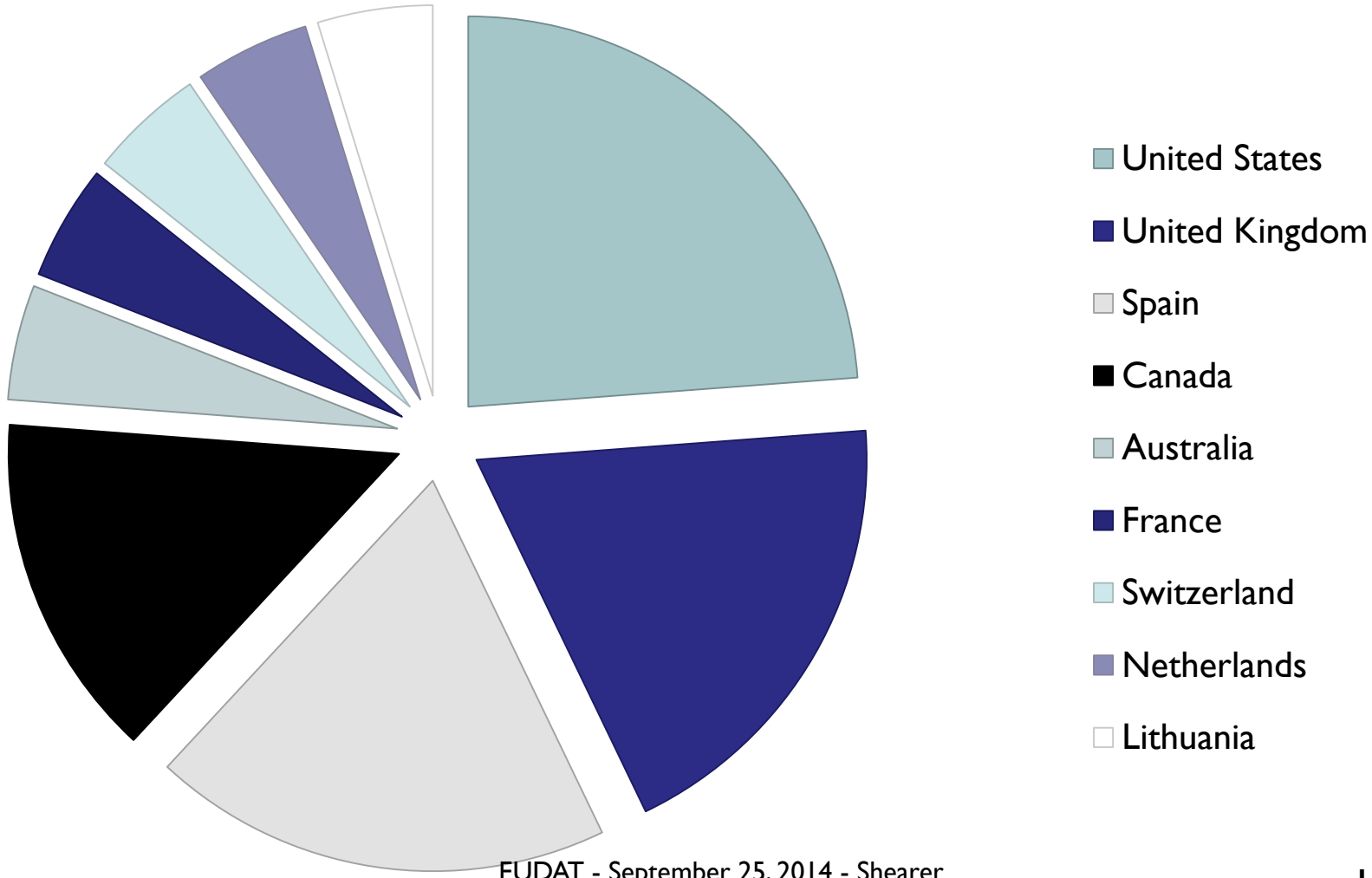
- Chuck Humphrey
- Stephen Marks
- Najla Rettberg
- Jochen Schirrwaggen
- Birgit Schmidt

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



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?

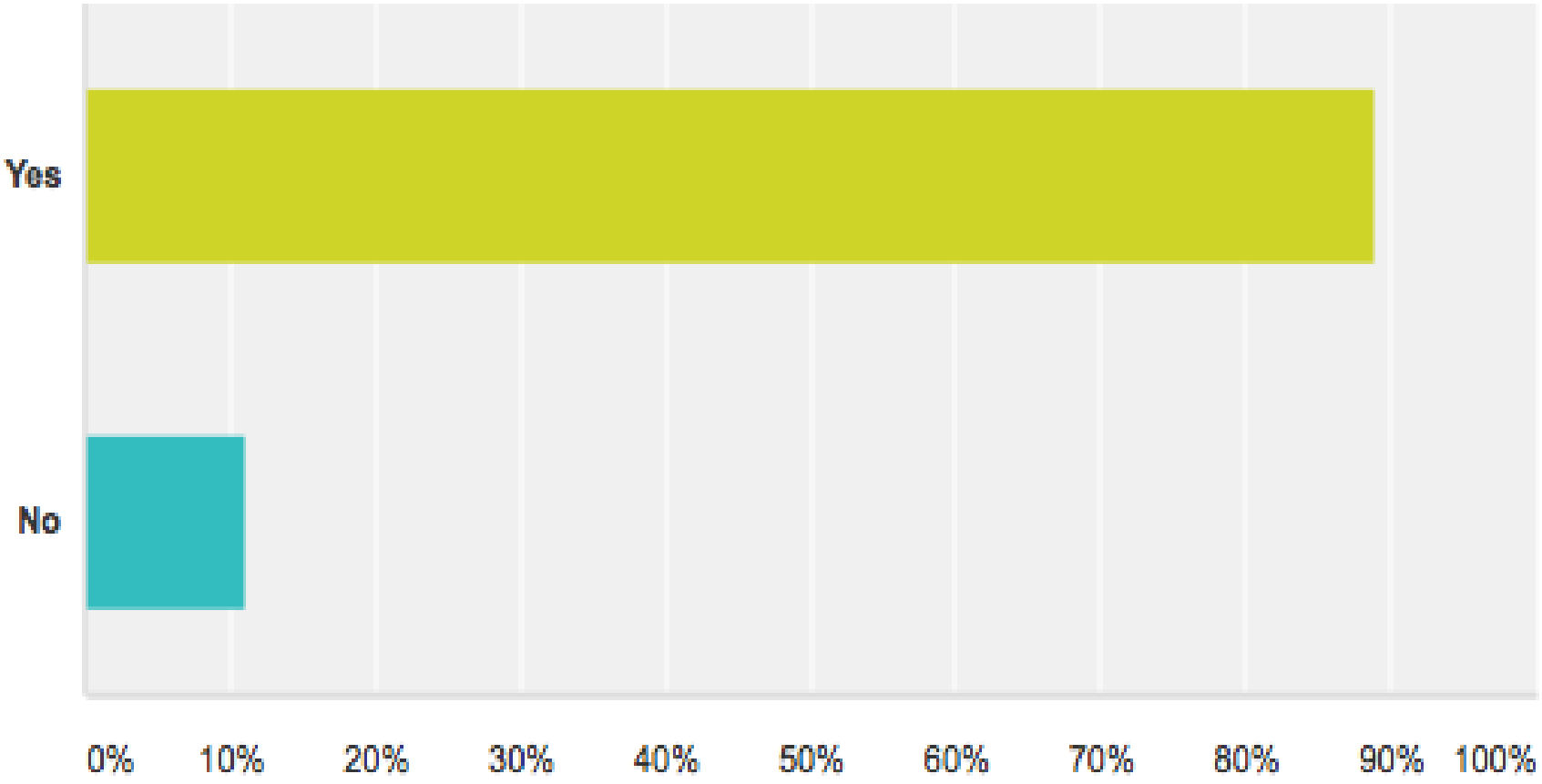
Repositories using a single schema

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…

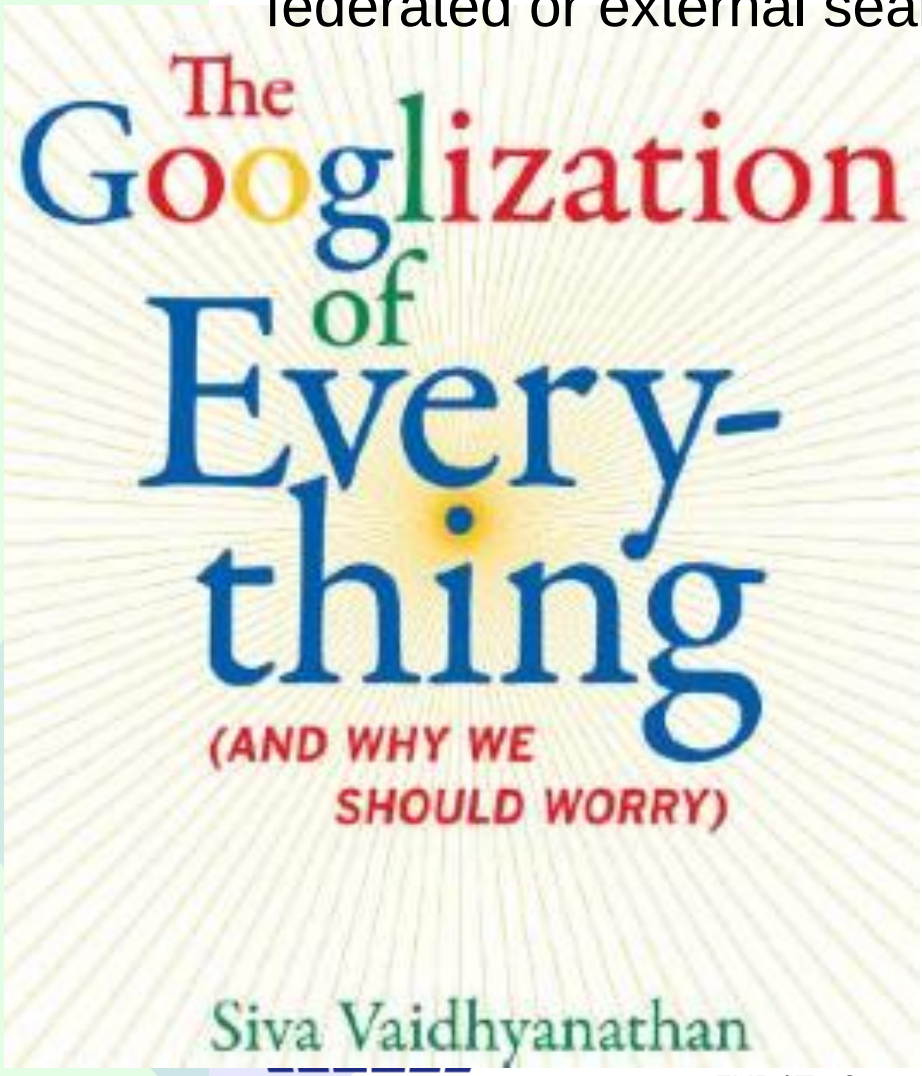
- The few datasets that we have appear to be easily discoverable in Google.
- Compliance with OAI-PMH ensures discoverability; also integration into library search schemes.
- By requiring that data to be associated with related publications, this enriches the metadata.
- We aim to index metadata to aid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