Wagging the Long Tail

Libraries and 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

Confederation of Open Access Repositories

- International association of repository initiatives
- Over 100 institutional members from around the world
- Vision: a global network of open access repositories in support of research and innovation

COAR Strategic Activities

Aligning repository networks

As research becomes increasingly global, it is critical to create infrastructure that can connect across geographic boundaries.

Tartu - October 23, 2014 - Shearer

Other Strategic Activities

Advocate for the "green road" and the institutional role in managing research outputs

Statement about embargo periods

Major international associations join together to underscore their support for immediate open access to research articles

As organizations committed to the principle that access to information advances discovery, accelerates innovation and improves education, we endorse the policies and practices that enable Open Access – immediate, barrier free access to and reuse of scholarly articles.

Pragmatic Activities

- Common vocabularies
- Usage metrics
- Linked data
- Impact and visibility of repositories
- Training and education
- And…

Research data!

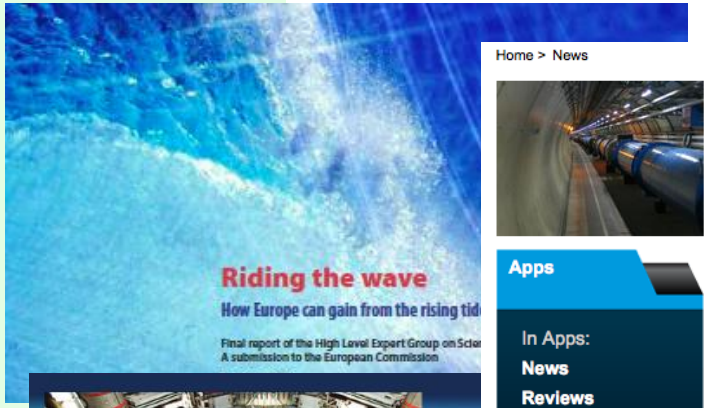
Our vision is a distributed network of data repositories (domain and institutional) that collect, manage and provide access to research data

- But this hinges on:

We don't want data silos!

- Long Tail for Research Data Interest Group
- Libraries for Research Data Interest Group (currently being reviewed by RDA)

“Big data” is all the rage!



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

Apps

In Apps:
News
Reviews
Features
How-tos
Slideshows

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.

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

Microsoft

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

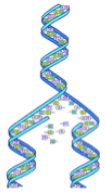
Jacqueline Vanacek, SAP

+ 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.”



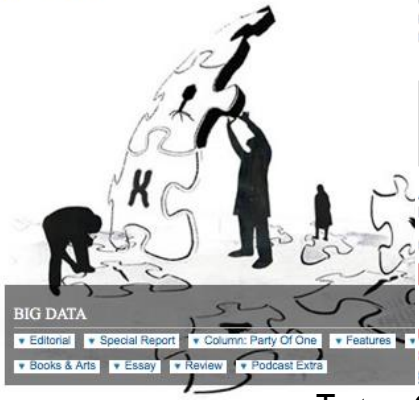
A slight mutation in matched nucleotides can lead to chromosomal aberrations and unintentional genetic

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Arc

BIG DATA // BIG DATA ANALYTICS

SPECIALS



NEWS
6/10/2014
07:06 AM

Jeff Bertolucci
News

Connect Directly

3 COMMENTS
COMMENT NOW
Login

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.

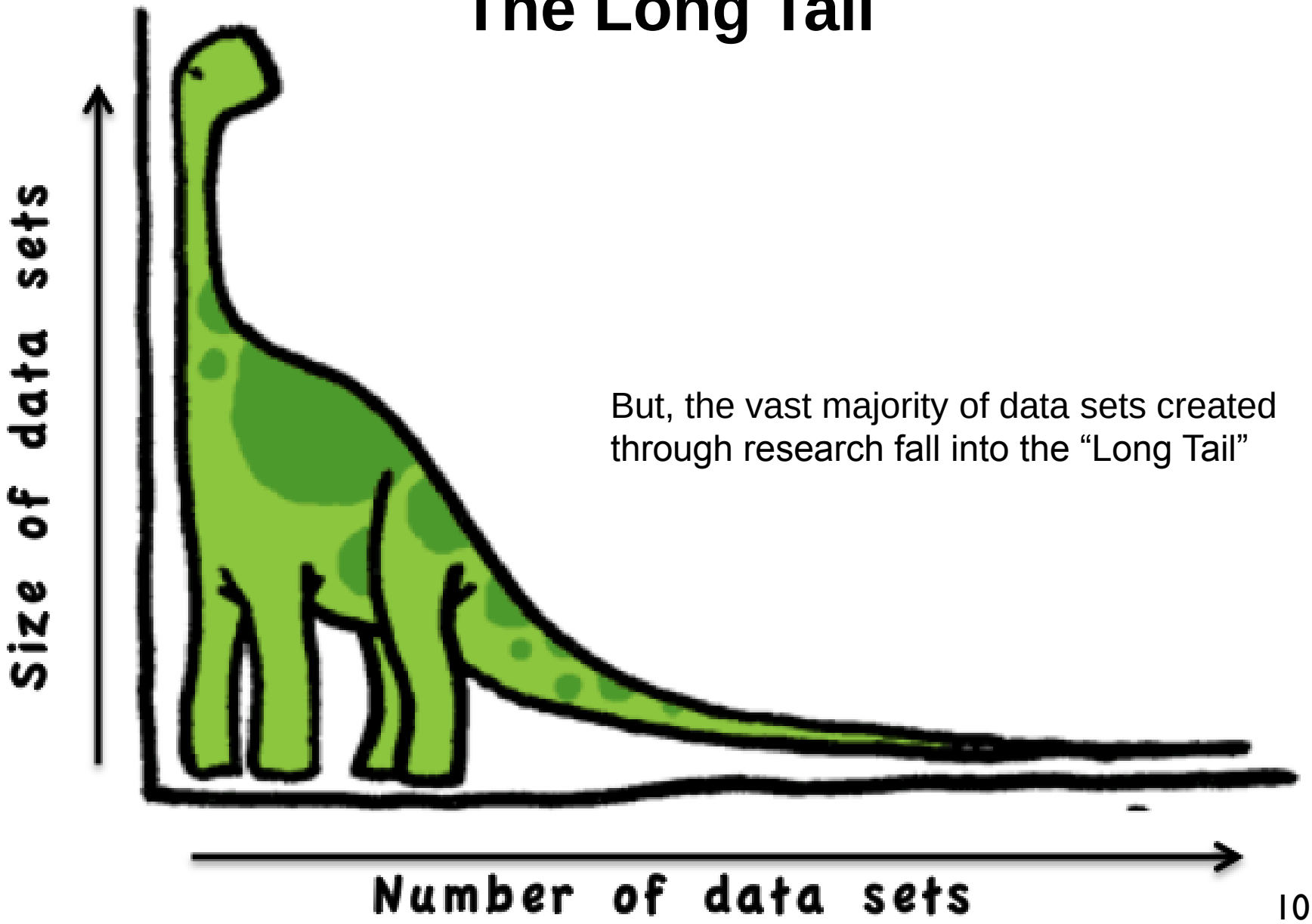
10 Big Data Pros To Follow On Twitter

(Click image for larger view and slideshow.)

EDITORIAL

video, mouseover for sound

The Long Tail



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)

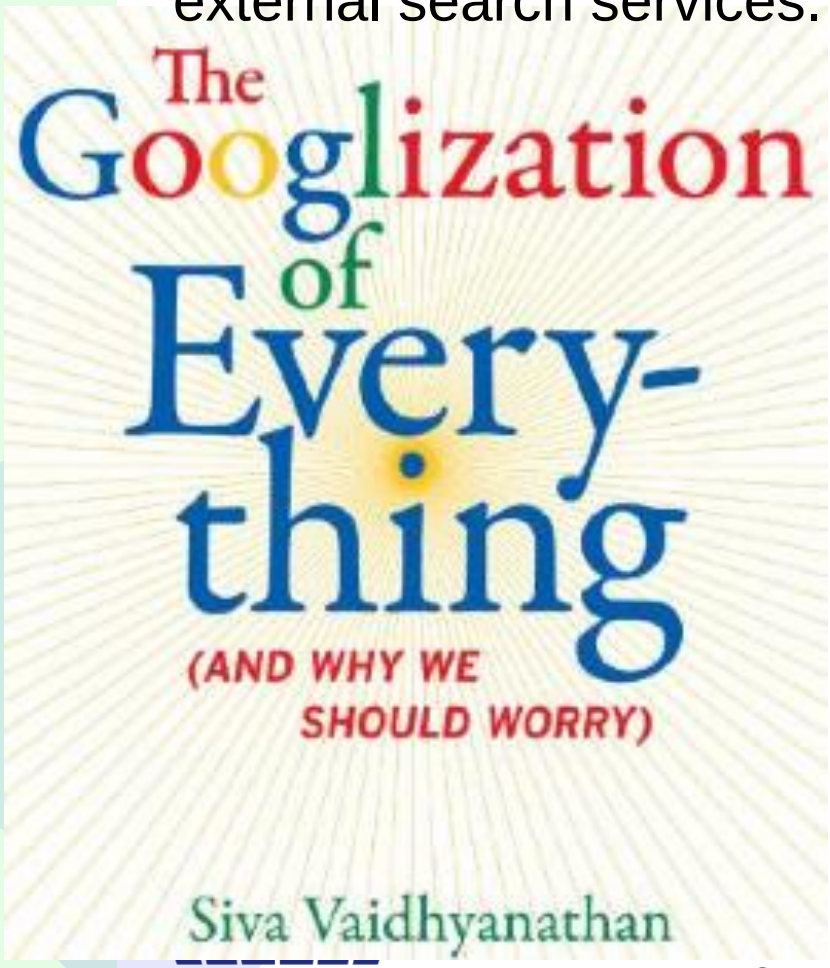
Metadata quality is a vital factor for electronic interoperability. (Rousidis, et al. 2014)

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

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

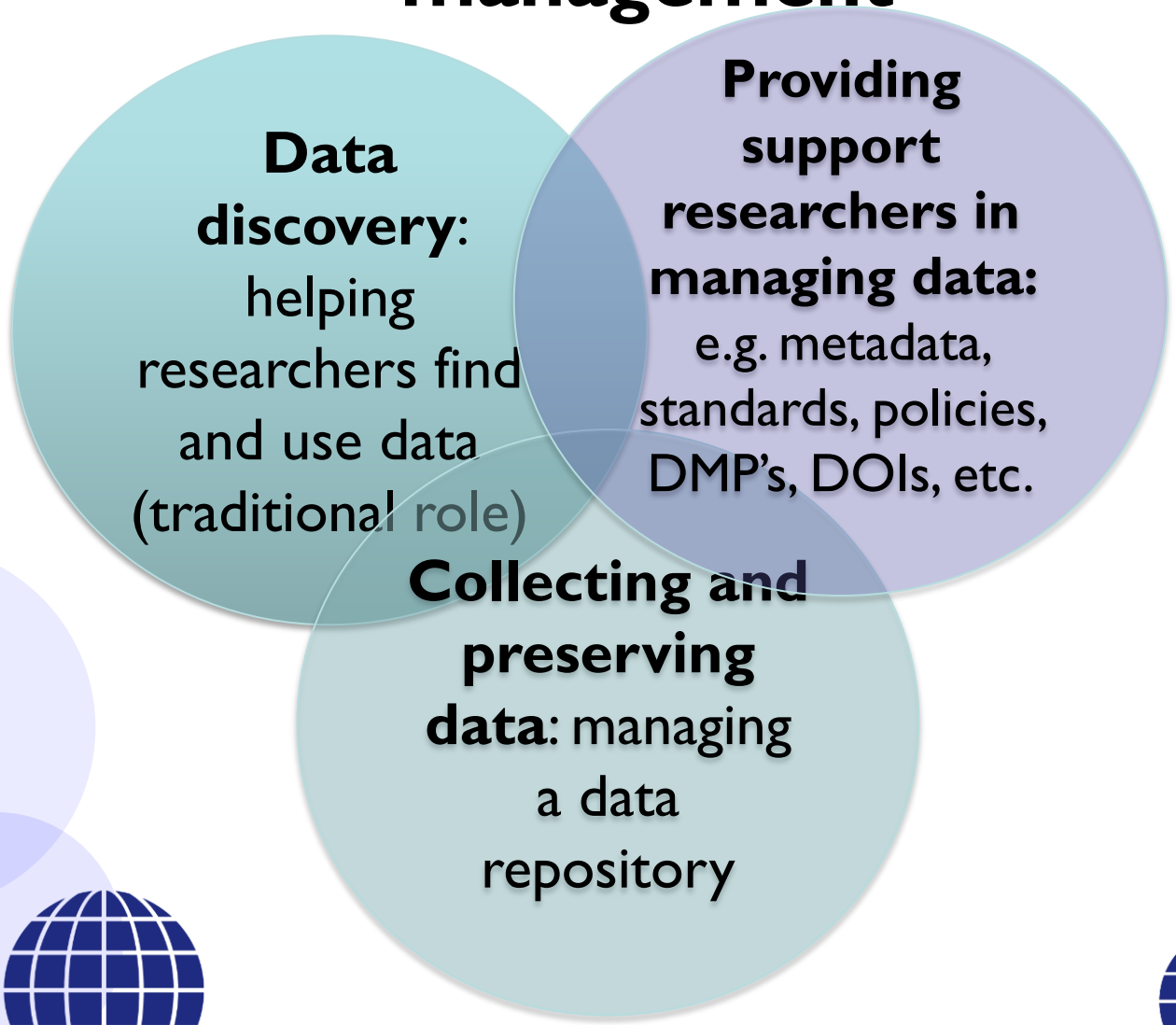


Yet, we know that most people use external services, such as Google as their main discovery tools.

Next steps for the RDA groups

- Incentives for deposit
- Identify key elements for interoperability across repositories and datasets
- Skills and training for data librarians
- Organizational models for library services in RDM

Library roles in research data management



Libraries and research data

Challenges:

- Blends new skills with traditional library expertise
- New organizational models
- Requires increased collaboration with other departments on campus (Information technology, researchers)
- Not universally accepted as falling in the scope of library services

Tänan! Questions?

Kathleen Shearer
Executive Director, COAR
kathleen.shearer@coar-repositories.org
www.coar-repositories.org