Long Tail Data Access

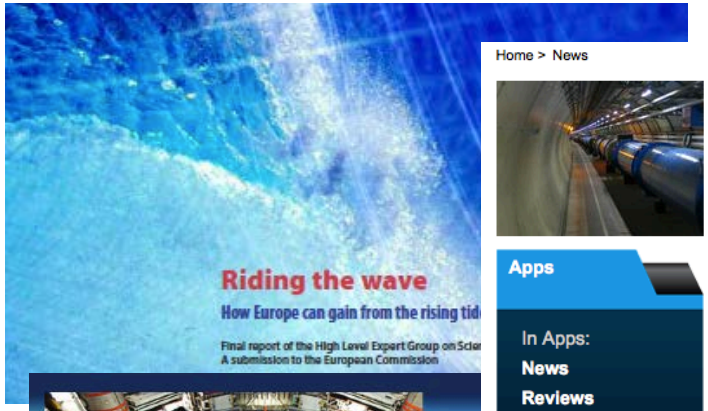
and the Research Data Alliance

Wolfram Horstmann

PASIG Karlsruhe, 17 September 2014

http://cdn.zmescience.com/wp-content/uploads/2009/08/jellyfish_nom.jpg

“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

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

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.

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

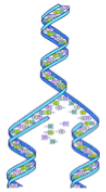
Jacqueline Vanacek, SAP

+ 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

SPECIALS

EDITORIAL

BIG DATA // BIG DATA ANALYTICS

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.)

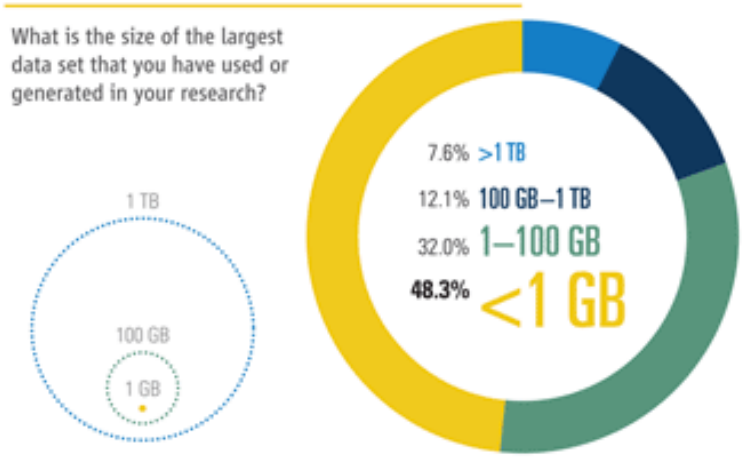
Sponsor video, mouseover for sound

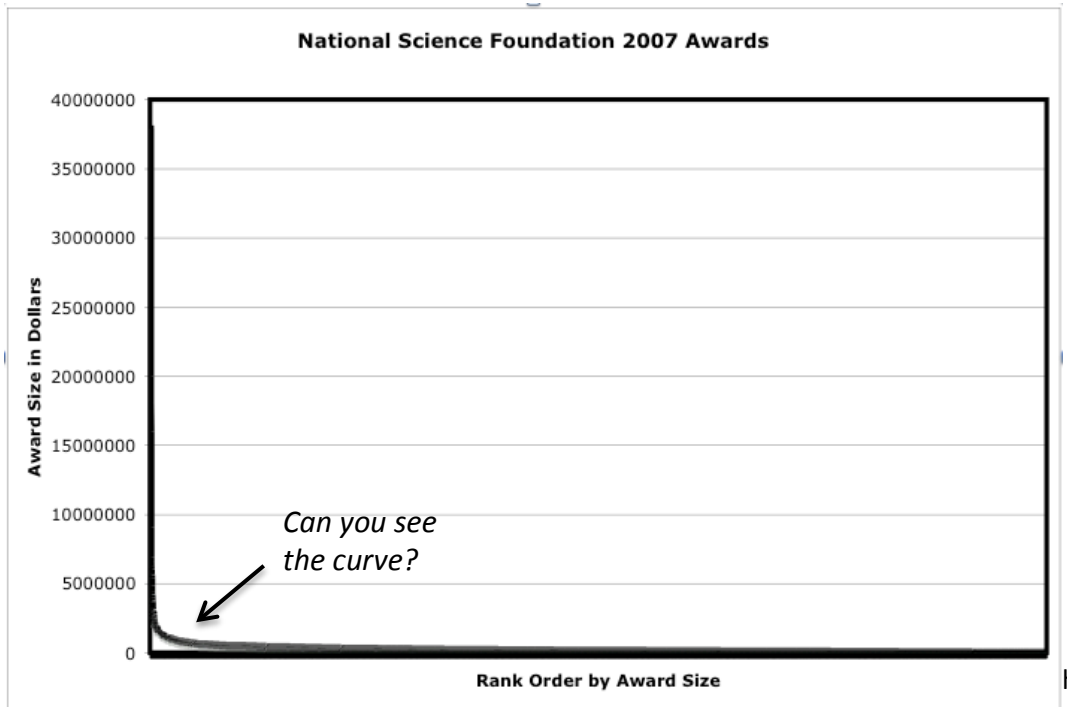
BUT

Larger parts of research use small data

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



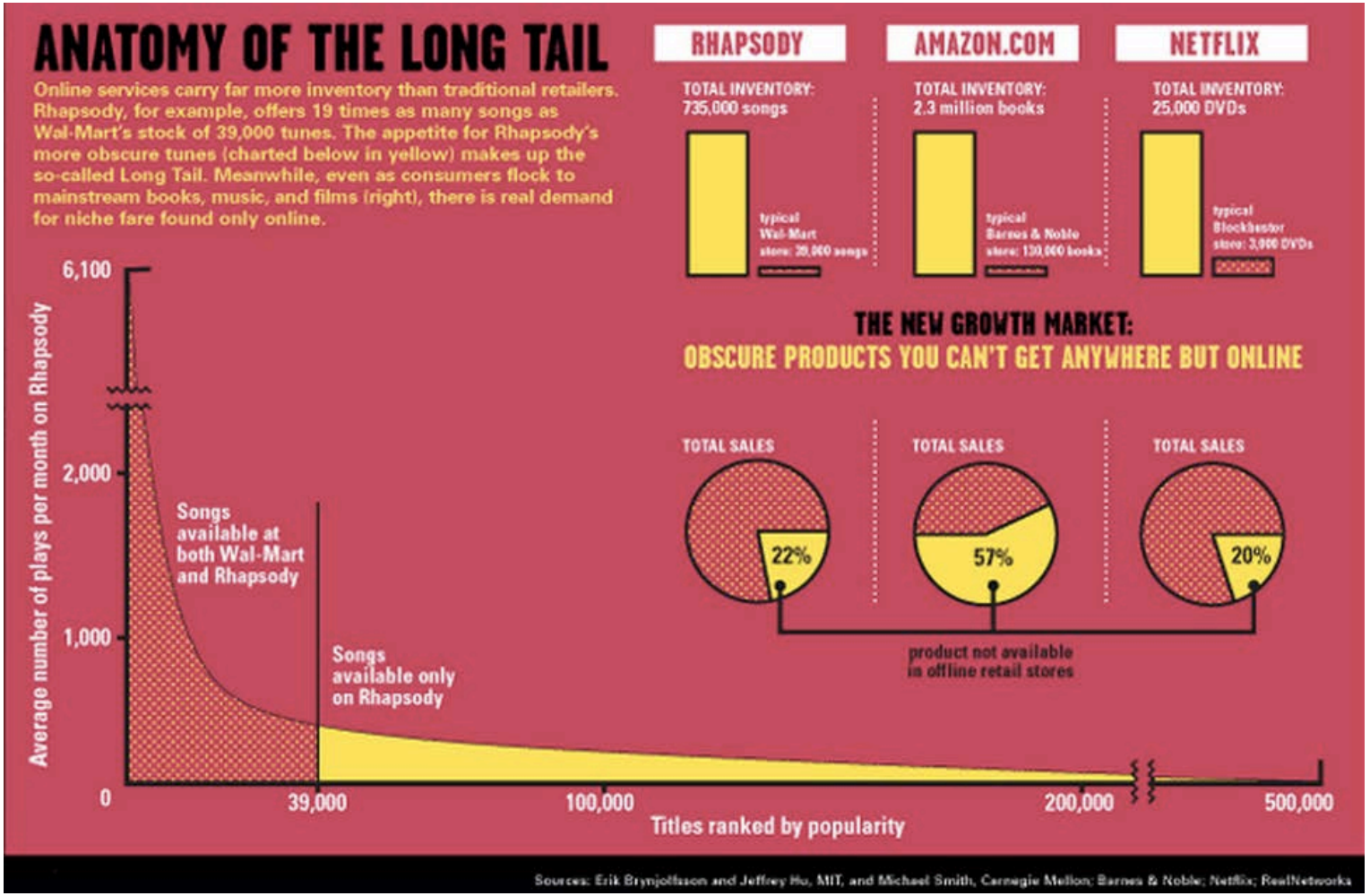


Because there is only a tiny fraction of large projects and a looooooooooooong tail of small projects

http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.pdf

“Long-Tail” as in Economics

Chris Anderson (Editor in Chief), Wired, Issue 12.10, October 2004



http://www.wired.com/wired/images.html?issue=12.10&topic=tail&img=2

“Long-Tail” as in Research Data

P. Bryan Heidorn (LIS U Arizona) in Library Trends 57/2, Fall 2008

- *… While great care is frequently devoted to the collection, preservation and reuse of data on very large projects, relatively little attention is given to the data that is being generated by the majority of scientists.*
- *… There may only be a few scientists worldwide that would want to see a particular boutique data set but there are many thousands of these data sets.*
- *… The long tail is a breeding ground for new ideas and never before attempted science.*
- *… The challenge for science policy is to develop institutions and practices such as institutional repositories, which make this data useful for society.*

http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.pdf

Big Data, Long-Tail 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

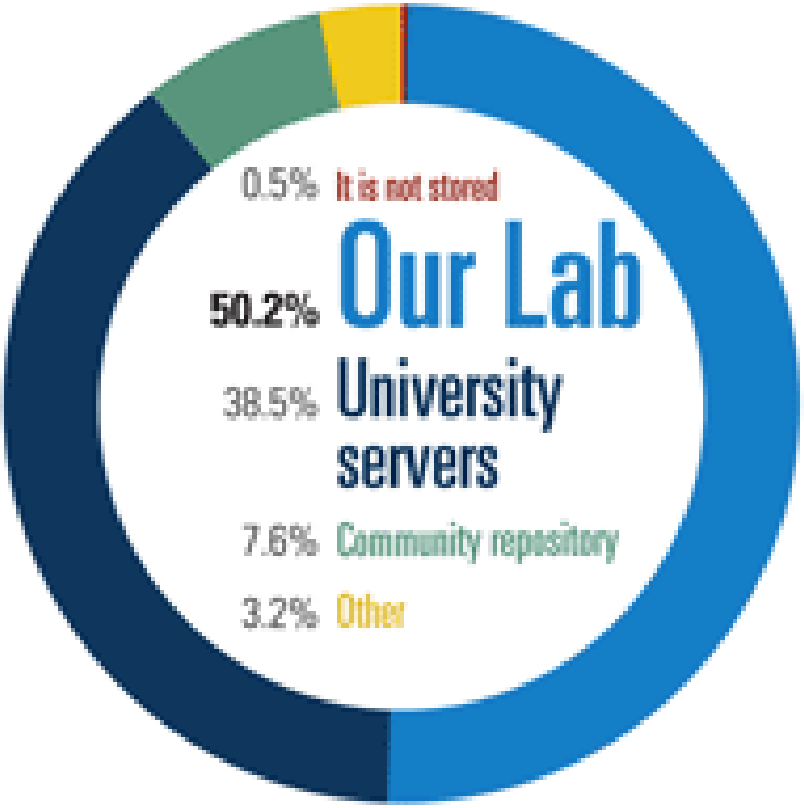
- "Disks in your drawer; server in lab basement"
- Long Tail Data exist across all disciplines

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

Institutional, domain or 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

Some of the challenges

Data quality

- appraise and show data as scientific / institutional /societal asset
- push standards for metadata and technology across disciplines

Discoverability

- increase discoverability in diverse repositories

Incentives

- show researchers how easy and beneficial it is to deposit data
- ask funders and institutions about policies

Business case

- show problems of irreproducibility, double research & innovation loss

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 challenges involved in managing diverse datasets
- To share and develop practices for managing diverse data
- To work towards greater interoperability across repositories

"Thanks for the slides, Kathleen!"

Kathleen Sheerer, COAR Executive Director and Co-Chair of the RDA IG

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
- make it easier for researchers to deposit their data
- sharing practices about discovery
- interoperability across repositories (WG!)
- preservation planning

Survey of Current Practices for Discovery Metadata

- Better understand practices with 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 representative but indicative

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…

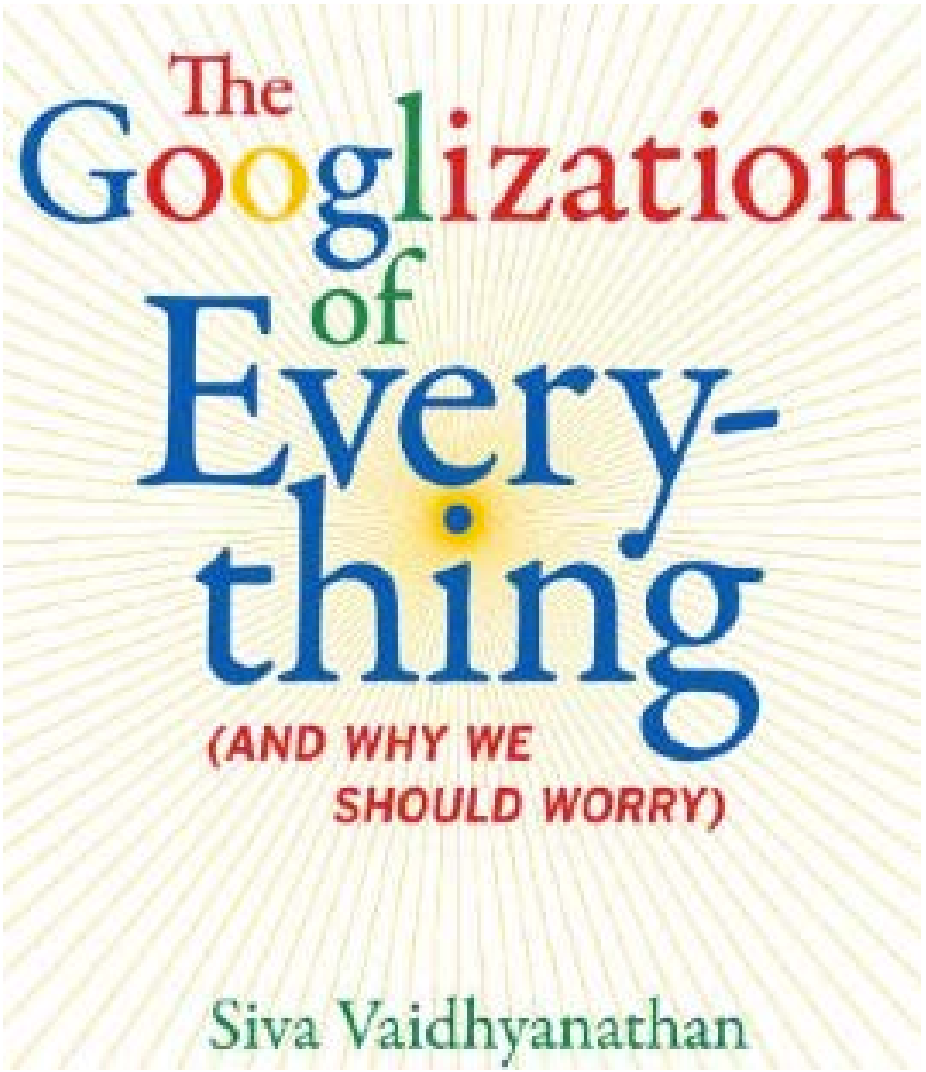
- Broadly speaking, and at a very high level, yes. If someone is looking for the data that supports a specific study, it is likely they will find it. However, if someone is looking for data with specific collection characteristics or other particularities then the metadata requires further enhancement.
- 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

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



Improving access to datasets

- Open licenses, funders and institutional policies
- Link data to publications, e.g. Force11, OpenAIRE
- Persistent Identifiers, e.g. DataCite-DOIs
- Discovery layer, e.g. landing pages, LOD, Schema.org
- Enable machine readability, e.g. APIs
- Dataset and repository registries, e.g. re3.org

Some Long-Tail Data Activities

- RDA IG Long-Tail Data
  - Meeting in Washington > www

- LIBER – Steering Group on Research Infrastructure
  - 10 recommendations and case studies > www

- COAR – Confederation of Open Access Repositories
  - Repository Interoperability Roadmap > www

- DRIVER/OpenAIRE
  - EU projects linking Literature to Data > www

Concluding remarks

1. Big research is based on small data
2. Reproducibility of large parts of research _?_
3. Institutions and funders should act
4. We must offer practical (!) solutions

THANKS