Beyond Big Data

The Long Tail of Research

Wolfram Horstmann

E-IRG, Riga, 3 June 2015
The mission I received

„Present main directions on Open Science along with main issues or areas that may need policy actions.“

Take Home Message

• Open Science needs to be on the ground and in the people, in 10.000s of labs around the world, in 100.000s of projects, performed by 1.000.000s of researchers.

• Big Data is only one – and maybe even a misleading – approach to implement Open Science

• The key to optimally exploit the capacity of Open Science lies in the Long Tail of Research
Some conceptual Questions

• Does infrastructure investment correlate with excellence in research?

• Does Big Data equal Big Science?

• Can you show me a Nobel Price based on Big Data?
“Big data” is all the rage

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

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.

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

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.”

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.

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.
Where is Big Data now?
Where is Big Data now?

Gartner ‘Hype Cycle’ Technologies 2014
http://www.gartner.com/newsroom/id/2819918
BIG DATA, LITTLE DATA, OPEN DATA
Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship.

(Christine Borgmann 2014)
The integrated scientific life cycle of embedded networked sensor research.

Figure 3. The integrated scientific life cycle of embedded networked sensor research.

(Pepe et al. 2009)
Beyond Big Data

...beyond the trendy discussion of 'big data' to focus on the real issue: data the very concept of which differs among scholarly communities...

J.L. King, University of Michigan on the book:


... the issue of sharing of the richly diverse and heterogeneous small data sets produced by individual neuroscientists, so-called long-tail data

“Long-Tail” as in Economics
Chris Anderson (Editor in Chief), Wired, Issue 12.10, October 2004

ANATOMY OF THE LONG TAIL
Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart’s stock of 39,000 tunes. The appetite for Rhapsody’s more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.

THE NEW GROWTH MARKET:
OBSCURE PRODUCTS YOU CAN’T GET ANYWHERE BUT ONLINE

Sources: Erik Brynjolfsson and Jeffrey Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; RealNetworks

“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
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.


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

http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.pdf
# Big Data, Long-Tail Data

<table>
<thead>
<tr>
<th>No.</th>
<th>Head</th>
<th>Tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Homogeneous</td>
<td>Heterogeneous</td>
</tr>
<tr>
<td>2</td>
<td>Large</td>
<td>Small</td>
</tr>
<tr>
<td>3</td>
<td>Common standards</td>
<td>Unique standards</td>
</tr>
<tr>
<td>4</td>
<td>Regulated</td>
<td>Not Regulated</td>
</tr>
<tr>
<td>5</td>
<td>Central curation</td>
<td>Individual curation</td>
</tr>
<tr>
<td>6</td>
<td>Disciplinary repositories</td>
<td>Institutional, general or no repositories</td>
</tr>
</tbody>
</table>

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
Heterogeneity

- 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/dsp/2013/06/14/arxiv-data-conservancy-pilot/](http://blogs.cornell.edu/dsp/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](http://wiki.datadryad.org/wg/dryad/images/b/b7/2013MayVision.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.”

- Our Lab (50.2%)
- University servers (38.5%)
- Community repository (7.6%)
- Other (3.2%)
- It is not stored (0.5%)

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
- 150 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
RDA Work, e.g. Discoverability

• 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. re3data.org
Some Long-Tail Data Activities

- **RDA Research Data Alliance, Long-Tail**
  - Regular Meetings >[www](#)

- **LIBER – Assoc. Europ. Reseach Libraries**
  - 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](#)

- **LERU – League of Europ. Research Universities**
  - Roadmap for Research Data >[www](#)
## Policy Actions for 'Long-Tail' Data

<table>
<thead>
<tr>
<th>No.</th>
<th>Long-Tail Characteristic</th>
<th>Do’s</th>
<th>Don’t’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Heterogeneous</td>
<td>Let a 1000 Flowers bloom and embrace diversity</td>
<td>Restrict innovation capacity by prescriptive normalisation</td>
</tr>
<tr>
<td>2</td>
<td>Small</td>
<td>Provide scale-to-size solutions</td>
<td>Enforce Big Data potpourri</td>
</tr>
<tr>
<td>3</td>
<td>Unique standards</td>
<td>Increase awareness and allow 'smart' standard customization</td>
<td>Ignore the context-specific expertise of data-management</td>
</tr>
<tr>
<td>4</td>
<td>Not regulated</td>
<td>Stress open licenses in publishing &amp; career incentives</td>
<td>Limit scientific and economic exploitation or personal rights</td>
</tr>
<tr>
<td>5</td>
<td>Individual curation</td>
<td>Support personalization and ownership transfer pipelines</td>
<td>Weaken the responsibility of data producers or researchers</td>
</tr>
<tr>
<td>6</td>
<td>Institutional, general or no repositories</td>
<td>Support interoperable institutional repositories in a global research data network</td>
<td>Fight distributed approaches and assume that only 'my' data centre knows how to manage data</td>
</tr>
</tbody>
</table>
Conclusions

• The Long Tail of Research – as I mean it – is the where the majority of innovation, citation, and data is generated
  – Simplification to Big Data and Big Science bears high risks of not using Europe’s research capacity

• Diversity is the main challenge but also the way research in organizing itself

• The institutional perspective needs to be fostered
--- and do not forget that the innovation is in the people...

THANK YOU