(Primary) Data: The New Special Collections for Research Libraries?

Wouter Schallier
Executive Director of LIBER (Association of European Research Libraries)

wouter.schallier@kb.nl
www.libereurope.eu

13th Fiesole Collection Development Retreat, St Petersburg, 11-13 May 2011
Contents

1. We live in a (primary) data world
2. E-science fiction?
3. Making (primary) data usable/useful
4. The new special collections for libraries?
5. Applause … and your questions
“COULD BECOME AS IMPORTANT A JOURNALISTIC TOOL AS THE FREEDOM OF INFORMATION ACT.” - TIME MAGAZINE

WikiLeaks is a non-profit media organization dedicated to bringing important news and information to the public. We provide an innovative, secure and anonymous way for independent sources around the world to leak information to our journalists. We publish material of ethical, political and historical significance while keeping the identity of our sources anonymous, thus providing a universal way for the revealing of suppressed and censored injustices.

WikiLeaks relies on its supporters in order to stay strong. Please keep us at the forefront of anti-censorship and support us today. You can also read more about WikiLeaks, our mission and objectives.

The Guantanamo Files: 779 classified prisoner dossiers revealed from the world’s most notorious prison

2011-04-28

In thousands of pages of documents dating from 2002 to early 2009 and never seen before by members of the public or the media, the cases of the majority of the prisoners held at Guantanamo — 758 out of 779 in total — are described in detail in memoranda from JTF-GTMO, the Joint Task Force at Guantanamo Bay, to US Southern Coand in Miami, Florida.
Open Government Initiative
TRANSPARENCY + PARTICIPATION + COLLABORATION

My Administration is committed to creating an unprecedented level of openness in Government. We will work together to ensure the public trust and establish a system of transparency, public participation, and collaboration. Openness will strengthen our democracy and promote efficiency and effectiveness in Government.

— PRESIDENT OBAMA, 01/21/09

13th Fiesole Collection Development Retreat, St Petersburg, 11-13 May 2011
It’s all about laaaaaaaarge amounts of data

“Data are no longer considered as interim products to be discarded once the research reporting them is published. Rather, they have become important sources of scholarly content to be used and re-used.”

Borgman, The role of libraries in e-science

Raw data from a central Pb+Pb event for 40 rows of the Main TPC
Image: http://na49info.web.cern.ch/na49info/Public/Press/pictures/mtpc40rowsRawData.gif
It’s about … (2)

- Sharing
- Use and re-use
- Using internet as a vehicle
- Distribution over many networked computers/people to analyse the data/make the data usable
- Transparency
- New expertise: data journalism, data managers, data scientists, data librarians etc.
The changing concept of publishing
E-science fiction?
Characteristics of e-science

- Large scale computing resources
- Data-intensive
- Carried out over the internet
- Collaborative (team science, virtual science communities)
- Distributed (networked science)
- Interdisciplinary
- Heterogeneous
The paradox of e-science

- I share my data because I want/need your data

Vs.

- I don’t share data because it doesn’t help my career
- It is MY data and I keep them safely stored on my laptop
- I don’t want other people to make misuse of my data
Opportunities for data exchange
http://www.ode-project.eu/

- Data sharing is
  - **smart**: it is efficient, avoids duplication, stimulates the advancement of science
  - about **transparency**: it allows re-analysis
  - about **enrichment**: it adds value to traditional publications
  - **rewarding**: requirement for publicly funded research
- 3 perspectives: researchers, publishers, libraries and data centres
Making primary usable/useful

- Data description and identification
- Organisation
- Data protection, privacy regulations, ethical issues
- Visualisation
- Interpretation
- Preservation
- Persistent link between publications and datasets
- Integrated search

http://na49info.web.cern.ch/na49info/Public/Press/LogBook.html
Making primary usable/useful (2)

- Validation and peer review of data
- Data quality and integrity
- Interoperability
- Repositories
- Control over correct usage
- Selection
- Data publication
- Citation
- …
Roles and responsibilities

- Researchers
- Students
- Publishers
- Data centres
- Libraries
- Funders: data management plans!

http://bookwormlibrarian.blogspot.com/2008/10/librarian.html
http://www.lgeoresearch.com/the-future-of-research/
Libraries and (primary) data

- Only 44% of responding institutions accepts research data for storage and preservation
- However, vast majority of data managers agrees that data preservation is important or very important for the following reasons:
  - Publicly funded research output should be properly preserved (98%)
  - Preserved data stimulates the advancement of science (96%)
  - It allows for re-analysis of existing data (95%)

2009 ParseInsight survey
Arguments AGAINST storing (primary) data in libraries

- Too many data
- Too complex
- Too expensive
- No money
- No people
- Many other challenges
Arguments IN FAVOUR of storing (primary) data in libraries

- There is no other way: e-science is a reality and primary data go hand in hand with it
- Who else will do it? Data needs stable and trustworthy storage

2009 No brief candle – reconceiving libraries for the 21st century
(Primary) data = new special collections?

- Often unique, difficult (expensive) to reproduce
- Intellectual capital of research institutes: success of research institutes will highly depend on their ability to generate and manage (primary) data, and to interpret, link and present their own and other’s data in an innovative way
(Primary) data = new special collections? (2)

- Libraries and data centres can directly contribute to more efficient/transparent research and education
- The success of libraries and data centres will depend on how successful they are in getting involved in the research and education workflows
- We need data managers! (Where are they?)
Examples

- http://datacite.org/
- http://datadryad.org/
- http://www.pangaea.de/
So the answer is YES!

- Thank you!

- Questions/comments/naughty remarks?

- Go home (or stay a while longer) and **get (primary) data**!