Research Data Management & Preservation: A Library Perspective

Brian Owen, Associate University Librarian – Library Technology Services & Special Collections, Simon Fraser University Library
LIBRARIES & BIG DATA

A Historical Perspective
Big Data Infrastructure circa 1925
Big Data Access Devices circa 1925
Big Data circa 1975
BIG DATA IN LIBRARIES CIRCA 1975

MARC RECORD

LDR04129:cam a220002538a 4500
001 103549699
003 SITKA
005 20110416203835.0
008 1006141s2011 nyu b 001 0 eng
010 1a 2010023588
020 1a 0231153074
020 1a 9780231153072
040 1a DLC 1c DLC
050 0 0 1a RJ240 1b .0385 2011
090 0 0 1a 615.372 032 2011
100 1 1a Offit, Paul A.
245 1 0 1a Vaccines and your child : 1b separating fact from fiction / 1c Paul A. Offit, Charlotte A. Moser.
260 1a New York : 1b Columbia University Press, 1c c2011.
300 1a 247 p.; 22 cm.
504 1a Includes bibliographical references and index.
1a Questions parents have about vaccines -- General -- What are vaccines? -- Why do we still need vaccines? -- How do vaccines work? -- How are vaccines made? -- What steps do pharmaceutical companies go through to make vaccines? -- Who recommends vaccines? -- How do we know vaccines work? -- Are vaccine-preventable diseases really that bad? -- Isn't it better to be naturally infected than immunized? -- Are vaccines given in a one-size-fits-all schedule? -- Is there any harm in using an alternative schedule? -- Why can't vaccines be combined to lessen the number of shots? -- Why aren't more vaccines given by mouth? -- Can I avoid vaccines by living a healthy lifestyle? -- Why should I trust a system that makes money for drug companies? -- Should vaccines be mandated? -- Is it my social responsibility to get vaccines? -- Safety -- Are vaccines safe? -- How do I know if a problem is caused by vaccines? -- What systems are in place to ensure that vaccines are safe? -- How do we know that different vaccines can be given at the same time? -- Do too
Big Data circa 2016

- 1000 Bytes = 1 Kilobyte
- 1000 Kilobytes = 1 Megabyte
- 1000 Megabytes = 1 Gigabyte
- 1000 Gigabytes = 1 Terabyte
- 1000 Terabytes = 1 Petabyte
- 1000 Petabytes = 1 Exabyte
- 1000 Exabytes = 1 Zettabyte
- 1000 Zettabytes = 1 Yottabyte
- 1000 Yottabytes = 1 Brontobyte
- 1000 Brontobytes = 1 Geopbyte
Big Data circa 2020?

• “Trends project a demand for 1-3 million Haswell-equivalent cores by 2020, and more than an exabyte of persistent storage. These projections may turn out to be underestimates, since some existing disciplines making extensive use of Compute Canada resources today anticipate needing over 1 million cores or 1 exabyte of data just for their own projects by 2020.” (Compute Canada Technology Briefing, November 12, 2015)
Library of Congress Holdings

- 32 million cataloged books and other print materials in 470 languages
- 61 million manuscripts
- over 1 million U.S. government publications
- 1 million issues of world newspapers
- 5.3 million maps
- 6 million works of sheet music
- 3 million sound recordings
- 4.7 million prints and photographic images including fine and popular art pieces and architectural drawings
“Every Six Hours, the NSA Gathers as Much Data as Is Stored in the Entire Library of Congress”

“There are 25 Petabytes \((10^{15})\) created every day and thrown into the internet. This is 70 times larger than the Library of Congress.”

“Facebook’s photo collection has a staggering 140 billion photos, that’s over 10,000 times larger than the Library of Congress.”

“Walmart handles 1 million plus customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data - the equivalent of 167 times the information contained in all the books in the Library of Congress.”
RESEARCH DATA MANAGEMENT & PRESERVATION

A Library Perspective
Stakeholders

• Compute Canada, CANARIE → Research Data Canada
• CFI, CIHR, NSERC, SSHRC
• CANFAR, CBRAIN
• U15, CUCCIO, CASRAI
• Institutional: VP Research, Research Ethics, IT Services
• CARL- Canadian Association of Research Libraries
• And don’t forget...researchers
RDM Spectrum

- **Canadian Genome Centres**
  - Currently 24 petabytes
  - Estimate 219 petabytes by 2020

- **Canadian Astronomy Data Centre**
  - 1.2 petabytes currently on Westgrid
  - Estimate 30-100 petabytes eventually needed

- **Ocean Networks Canada (Neptune, Venus)**
  - 300+ terabytes of data archived
  - 170 gigabytes of data collected every day
  - 230 gigabytes of data are distributed every day

- **Pacific Herring Project**
  - 26 narrative videos totalling about 5 megabytes
Data Curation

- Digital curation is maintaining and adding value to a trusted body of digital research data for current and future use; it encompasses the active management of data throughout the research lifecycle.
Research Data Management Issues

- data integrity
- preservation
- discovery, access and authentication
- re-use
- policies and procedures (local, national, funding agencies; also intellectual property)
- inter-operability and participation in larger national and international initiatives
- standards and metadata
• Developing a national research data culture
• Fostering a community of practice for research data
• Building national research data services and infrastructure
• **DMP Assistant** is a bilingual tool for preparing data management plans (DMPs). The tool follows best practices in data stewardship and walks researchers step-by-step through key questions about data management.

• [https://assistant.portagenetwork.ca/](https://assistant.portagenetwork.ca/)
Discovery and Preservation

• Library-based projects (2012-15)
  o SFU: Research data repository & preservation
  o UBC: Research data repository & preservation
  o Univ. of Alberta: Canadian Polar Data Network
  o OCUL & Scholars Portal: Repository & Cloud Storage

• Outcomes
  o Prototypes and limited production systems
  o Experience
Discovery and Preservation

• Research Data Canada Federated Pilot (2014-15)
• Participants: Compute Canada, CANARIE, CARL
• Software Platforms:
  o Dataverse/Islandora (Discovery)
  o Archivematica (Preservation)
  o Globus (Replication & Discussion)
• Libraries: repository and preservation tools, metadata standards, researcher needs
• Compute Canada: replication tools, robustness and scalability assessment
• Outcomes
  o More experience
CARL & COMPUTE CANADA MOU (2016)

• CARL/Portage:
  o research data management expertise,
  o metadata and workflow design / implementation
  o contributing to requirement specifications and design
  o obtaining research data for test purposes
  o liaising with researchers, repository services, and preservation software developers

• Compute Canada:
  o software development and computational resources,
  o Globus liaison
  o for liaising with Globus
  o costs of conducting these activities
Conclusions

• “state of churn”
  o Storage and infrastructure requirements
  o RDM platforms, tools and workflows
  o Policies and procedures
  o Roles and responsibilities
  o Costs to sustain
Thank You

Questions?

Brian Owen
brian_owen@sfu.ca