

Jun 23rd, 8:45 AM - 10:00 AM

## Data curation: just in time, or just in case?

Michael Lesk  
*Rutgers University*, [lesk@rci.rutgers.edu](mailto:lesk@rci.rutgers.edu)

Follow this and additional works at: <http://docs.lib.purdue.edu/iatul2010>

---

Michael Lesk, "Data curation: just in time, or just in case?" (June 23, 2010). *International Association of Scientific and Technological University Libraries, 31st Annual Conference*. Paper 5.  
<http://docs.lib.purdue.edu/iatul2010/conf/day3/5>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact [epubs@purdue.edu](mailto:epubs@purdue.edu) for additional information.

# Data curation: just in time, or just in case?

*The best is the enemy of the good* – Voltaire

Libraries have an opportunity in data storage; but we can't afford good curation of all data from the start. We need selection, delay, cooperation and tools; and we need new social and economic models.

# Data

The internet breaks business models; it's also broken the academic science model.

Open data accelerates science – look at astronomy or molecular biology.

As Jim Gray and Chris Anderson both pointed out, data collection is being separated from hypothesis evaluation. But... if data is collected ahead of when it is needed, how do we know how much we need to keep? And how can we afford to do that?

Economics, culture, and technology all collide.

# Motivate the researchers

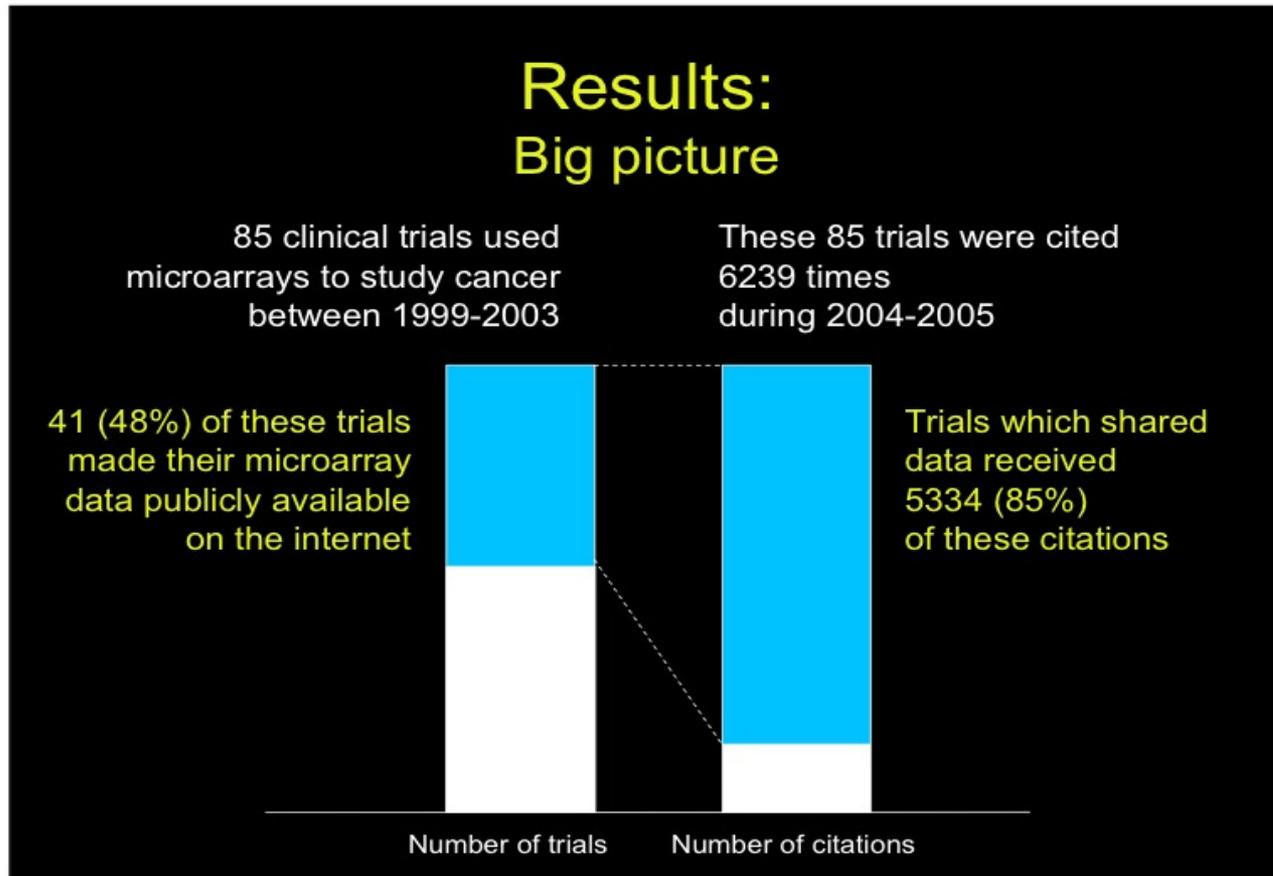
The problem – as a quote from a blogger praising Helen Berman:

*One of the remarkable things about Helen is that her life has been devoted to service within science rather than, as some might call it, doing real science. By concentrating on the infrastructure her contribution, I believe, is much greater than if she had just run her own lab, even a very successful one.*

The first sentence infuriates me. (Helen Berman has led the Protein Data Bank for decades).

Tenure is a big issue: this might really be easier in an engineering school or department.

# Piwowar, *Nature Precedings*, 2007



And, indeed, the 48% of trials which made their data available received 85% of the total citations... clearly more than their fare share.

# What's the fear?

Optimist: fear of being scooped with results on your own data

Pessimist: fear that your work will be shown to be wrong.

The latter is not a joke. B. D. McCullough, “Got Replicability? The Journal of Money, Credit and Banking Archive”: *Econ Journal Watch*, v. 4, pp 326-337, 2007: of 193 articles that should have had deposited data, only 14 could actually be replicated. See also Savage & Vickers in *PLoSOne*. And when work is replicated, half the time the results are different: see Hubbard and Vetter, “An empirical comparison of published replication research in accounting, economics, finance, management, and marketing” *J. Business Research*, v. 35, pp. 153-165, February 1996

# What to do?

Social research:

Why the differences across fields?

Why the differences across universities?

Scooping:

embargo periods,

a journal-imposed rule requiring citation to the data.

Errors:

if you know others will see your data you'll be more careful?

Choosing the patrons: retiring faculty, new PhDs both mentioned here.

# Economics

Curation is expensive: one study at Oxford suggested that curation costs about 1/3 as much as the original experiment does! (Ditto for data handling costs in astronomy and ocean science).

We'd better hope these numbers are wrong.

They are all costs incurred up front – before the data is used, before we even know if it will be useful.

Some of this could be pushed back to the users if they either recognize the value of the curation or are sufficiently pressured by NSF or NIH.

But libraries should beware of demands for “quality.” We would not have either Google Books or the Open Content Alliance at the per-page price BL spent to scan *Beowulf*.

# Any source of money?

We heard some complaints yesterday about the cost of open access journals. Yes, Elsevier complains that PLoS is only recovering half its costs at \$1500/article. Well, Elsevier publishes about 250,000 articles a year and has revenues of \$2.5B. Sounds like about \$10,000 per article (you can make allowances for rejected articles and for their books and other products).

ARL libraries spend about \$800M/yr on journals; just like data, a lot of money is spent up front for items that might not be used. This is going to fall apart, somehow.

# Disciplinary vs. institutional archives

It doesn't seem realistic to think that every university will have a data specialist in every area.

Yet the disciplinary archives are mostly soft-money; the UK is a disturbing story of money for them being removed.

Can we do this with cooperation? Locally managed but with expertise outsourced (and quite possibly the storage in a kind of cloud, whether Amazon or LOCKSS)?

There is work to be shared within an institution: user assistance, data migration, ...

# Do-it-yourself?

Could we ask people to do the deposit themselves with volunteer monitoring and checking?

Obvious risk: with no refereeing or formal curation, lots of trash gets deposited (see earlier references to economics archives).

But: arXiv seems to work; Wikipedia seems to work; “deferred examination” patenting seems to work.

And with time, tools will get better. Bill Arms suggested this a long time ago.

# Platitudes

Anything worth doing is worth doing well.

Anything worth doing is worth doing badly.

Both true.

And it's not as if the traditional system is all that perfect;  
human cloning, cold fusion or polywater anyone?

What's being used will be given attention faster – before those  
who collected it have retired.

# Citizen science

Beyond asking faculty or students to do data deposit, look at some of the completely volunteer efforts:

Project Feederwatch?

Globe?

Enhancement of astronomy databases?

Distributed Proofreaders?

And some almost-professional: Dave Bertelsen has hiked the same desert trail 1270 times since 1981 and made 195,000 observations of the biota (summer 2010, *OnEarth* magazine).

# Better tools?

Compare Google transit with the Text Encoding Initiative; about 10 pages vs. over 1400 pages.

Data fusion: have been looking with J. Gelernter at aligning survey questions in ICPSR data.

Look at the Internet Archive, which is saving more bytes per dollar than anyone else – minimal metadata, what there is comes from volunteers.

I'm ignoring  
confidentiality,  
privacy, IP rights...

a lot of scientific  
data doesn't involve  
these problems.



**YOU ALREADY HAVE  
ZERO PRIVACY.  
GET OVER IT.**

Scott MacNealy, CEO, Sun Microsystems

# Fear and greed

The astronomers have > 100 TB of data, and 0.5 TB of journal articles. They don't need help with the articles if they store the data. But then what's left for libraries?

There really is a need for expertise in selection, long-term preservation, and helping users. Libraries are good at those things.

But we need to focus on *good enough*, on *when needed*, and on *getting help*.