Prevalence and Patterns of Biomedical Research Data Sharing and Reuse

Heather Piwowar
Department of Biomedical Informatics
University of Pittsburgh

JCDL Doctoral Consortium
June 2008

$$

?

Is it working? Is it worth it?

Are scientists sharing their data?

Are other scientists reusing the data?

Who, if anyone, is benefiting from the policies, tools, and initiatives?

We cannot manage
what we do not measure

Dissertation Objective:

Evaluate the patterns and prevalence of biomedical research data sharing and reuse

Prior work in this area

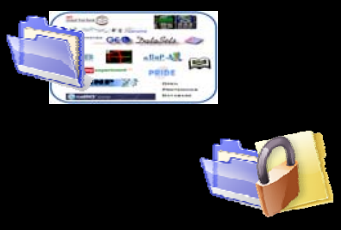
- Surveys
- Manual audits
- Automated classification of citation contexts

See http://www.citeulike.org/user/hpiwowar for bibliography

Missing:
a study of data sharing and reuse behavior and impact
based on a broad spectrum of instances

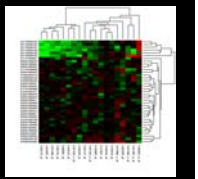
Dissertation Research Questions

1. prevalence of sharing and reuse
2. patterns of sharing and reuse
3. affect of sharing and reuse on impact
4. implications of findings



Data type

- Gene expression microarrays



http://en.wikipedia.org/wiki/DNA_microarray
http://en.wikipedia.org/wiki/Image:Heatmap.png

Sharing type

- Openly online, mentioned in publication
  - PubMed
  - filtered with MeSH terms for gene-expression
  - English, machine-readable full text
  - 2000-2007

Dissertation dataset

Article ID	Link to full text			
234	http://…			
456	http://…			
657	http://…			
897	http://…			

1.What is the **prevalence** of biomedical research data sharing? of biomedical research data reuse?

Endpoints

- Three endpoints:
  - Does this study produce raw data?
  - If so, does this study share the raw data?
  - Does this study reuse others' raw data?

Example text cues

- Sharing
  - "our data has been deposited in the GEO database"
  - "the microarray expression values from this study are available at the following website"
- Reuse
  - "using the data of Smith et al, we… "
  - "we downloaded four datasets from…"

Identification of endpoints

- Train and evaluate a Natural Language Processing system to recognize endpoint cues within full text
- Performance to be summarized as precision and recall with confidence intervals

Pilot data for NLP identification of sharing: Piwowar and Chapman, submitted to AMIA 2008.

Research Question 1: Prevalence

		Endpoints		
Article ID	Link to full text	Produces data?	Shares data?	Reuses data?
234	http://…	TRUE	TRUE	FALSE
456	http://…	TRUE	TRUE	TRUE
657	http://…	TRUE	FALSE	FALSE
897	http://…	FALSE	n/a	TRUE

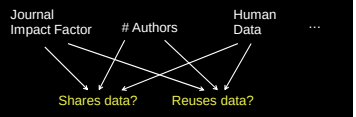
Calculate Percentages

2. What **features** are most associated with an investigator's decision to share or reuse a biomedical research dataset?

- Features to include:
  - Journal Impact Factor
  - Number of Authors
  - Are the samples from humans
  - Subdiscipline
  - Strictness of Journal policy on data sharing
  - Institution
  - Year of publication
  - …

Pilot data for journal policies: Piwowar and Chapman, ELPUB 2008.

Endpoints		Covariates					
Article ID	Link to full text	Produces data?	Shares data?	Reuses data?	Journal Impact Factor	# Authors	Human Data
234	http://…	TRUE	TRUE	FALSE	1.5	2	TRUE
456	http://…	TRUE	TRUE	TRUE	23.5	1	FALSE
657	http://…	TRUE	FALSE	FALSE	2.4	6	FALSE
897	http://…	FALSE	n/a	TRUE	0.6	2	TRUE

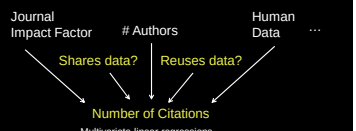


Multivariate logistic regressions

3. Does sharing or reusing data contribute to the **impact** of a research article, independently of other factors?

Assumption:
citation count is a proxy for research impact

Endpoints					Covariates			
Article ID	…	Produces data?	Shares data?	Reuses data?	…	…	…	Number of Citations
234		TRUE	TRUE	FALSE	1.5	2	TRUE	0
456		TRUE	TRUE	TRUE	23.5	1	FALSE	4
657		TRUE	FALSE	FALSE	2.4	6	FALSE	4
897		FALSE	n/a	TRUE	0.6	2	TRUE	5



Multivariate linear regressions

Pilot data on citation impact for sharing: Piwowar, Day and Fridsma, PLoS ONE 2007.

4. What do the results **suggest** for developing efficient, effective policies, tools, and initiatives for promoting data sharing and reuse?

We might discover, for example:
- Lots of sharing for non-human data
- All reuse within the first 5 years
- Journal and funder requests are ineffective

Significance

- Dataset
  - social network analysis, simulation, …
- NLP Classifiers
- Best-practice patterns, communities
- Novel research connections
- Inspire further work in this area
  - policy evaluation, reusability metrics
  - citations for data (Data Reuse Registry)

Piwowar, Chapman. Envisioning a Data Reuse Registry. Poster submitted to AMIA 2008.

Limitations

- Causation?
- Other data types?
- Other sharing mechanisms?

"Does anyone want *your* data?

That's hard to predict ¿-¡
After all, no one ever knocked on your door asking to buy those figurines collecting dust in your cabinet before you listed them on eBay.

Your data, too, may simply be awaiting an effective matchmaker."

Got data? Nature Neuroscience 10, 931 (2007)

My data is here

www.dbmi.pitt.edu/piwowar

I urge you to share yours, too.

Thank you

Funding: *NLM informatics training grant*
Advisor: *Dr. Wendy Chapman*
Committee: *Dr. Ellen Detlefsen*
Dr. Madhavi Ganapathiraju
+ JCDL Funding and Reviewers!

Questions, Comments, or Suggestions?