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

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. Effect of sharing and reuse on impact
4. Implications of findings

Data type
• Gene expression microarrays
Sharing type
• Openly online, mentioned in publication
  – PubMed
  – filtered with MeSH terms for gene-expression
  – English, machine-readable full text
  – 2000-2007

Dissertation dataset

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

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

Research Question 1: Prevalence

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

<table>
<thead>
<tr>
<th>Article ID</th>
<th>Shares data?</th>
<th>Reuses data?</th>
<th>Publications type</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>456</td>
<td>TRUE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>657</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>897</td>
<td>FALSE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Article ID</th>
<th>Shares data?</th>
<th>Reuses data?</th>
<th>Publications type</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>456</td>
<td>TRUE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>657</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>897</td>
<td>FALSE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Article ID</th>
<th>Shares data?</th>
<th>Reuses data?</th>
<th>Publications type</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>456</td>
<td>TRUE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>657</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>897</td>
<td>FALSE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Article ID</th>
<th>Shares data?</th>
<th>Reuses data?</th>
<th>Publications type</th>
</tr>
</thead>
<tbody>
<tr>
<td>234</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>456</td>
<td>TRUE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>657</td>
<td>TRUE</td>
<td>FALSE</td>
<td>2000-2007</td>
</tr>
<tr>
<td>897</td>
<td>FALSE</td>
<td>TRUE</td>
<td>2000-2007</td>
</tr>
</tbody>
</table>

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
– …
Article ID

Link to full text

Produces data?

Shares data?

Reuses data?

Journal Impact Factor

Authors

Human Data

Number of Citations

Endpoints

Endpoints

Endpoints

Endpoints

Number of Citations

Number of Citations

Number of Citations

Number of Citations

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

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

Figure: Multivariate logistic regressions

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)

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.”

Figure: Multivariate linear regressions

Figure: Multivariate linear regressions

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 Deetjen
Dr. Madhavi Ganapathiraju
+ JCDL Funding and Reviewers!

Questions, Comments, or Suggestions?