Open for all. Reusable for whom?

A review of what data reusers want and how data repositories can deliver
Motivation for Our Work

Examples of prior data reuse research:

Curty (2016) provides an integrated view of the individual, social, and technological factors that influence reuse.

Yakel et al., (2013) describe how reusers construct trust in repositories and how they associate repository actions with trust.

Chin and Lansing (2004) design the Biological Sciences Collaboratory after identifying and capturing social and scientific contexts needed to understand and reuse others’ data.

Our research questions:

How do reusers needs for context translate into a desirable feature set for data repositories?

How do data repositories showcase these desirable features in use?
Context from the Data Reuser’s Point of View

Repository Information
- Provenance
- Reputation and History
- Curation and Digitization

Data Production Information
- Data Collection
- Specimen and Artifact
- Data Producer
- Data Analysis
- Missing Data
- Research Objectives

Data Reuse Information
- Prior Reuse
- Terms of Use
- Advice on Reuse

(Faniel, Frank, & Yakel 2019)
Context Types by Percentage of Mentions

Data Collection
Provenance
Specimen and Artifact
Data Producer
Data Analysis
Repository Reputation and History
Prior Reuse
Curation and Digitization
Missing Data
Advice on Reuse
Research Objectives
Terms of Use

% of social scientists (n=43)  % of zoologists (n=40)  % of archaeologists (n=22)

(Faniel, Frank, & Yakel 2019)
Our Approach

➢ Translated the twelve context types into a desirable feature set
➢ Reviewed generalist and disciplinary data repositories for those features
➢ Created a self-assessment questionnaire for data repository staff asking what context they capture and display
➢ Piloted the self-assessment questionnaire with staff at four data repositories in the Data Curation Network (DCN)
➢ Conducted interviews with staff to discuss self-assessment results, including what they identified as their data repositories’ exemplar features.
<table>
<thead>
<tr>
<th>Categories</th>
<th>Data Repository Features Supporting Reuse Context</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data production</td>
<td>has no categories.</td>
<td></td>
</tr>
<tr>
<td>information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Collection</td>
<td>Methodology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ephemeral documentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data type</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Granularity or resolution</td>
<td></td>
</tr>
<tr>
<td>Specimen and Artifact</td>
<td>Taxonomies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provenance/source</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unique IDs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Access info</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Age estimation</td>
<td></td>
</tr>
<tr>
<td>Data Producer</td>
<td>ORCID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other Data Producer</td>
<td></td>
</tr>
<tr>
<td>Data Analysis</td>
<td>Code book</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation of data analysis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reproducibility</td>
<td></td>
</tr>
<tr>
<td>Missing Data</td>
<td>Documentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Procedure</td>
<td></td>
</tr>
<tr>
<td>Research Objectives</td>
<td>Linkages research</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linkages funder</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Abstract</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data citations</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Categories</th>
<th>Data Repository Features Supporting Reuse Context</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provenance</td>
<td>By source</td>
<td></td>
</tr>
<tr>
<td></td>
<td>By traceability</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Versions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linkages</td>
<td></td>
</tr>
<tr>
<td></td>
<td>File naming convention</td>
<td></td>
</tr>
<tr>
<td>Repository</td>
<td>Support services</td>
<td></td>
</tr>
<tr>
<td>Reputation and History</td>
<td>Staffing levels</td>
<td></td>
</tr>
<tr>
<td>Curation and Digitization</td>
<td>Documentation about the digitalization and curation process.</td>
<td></td>
</tr>
<tr>
<td>Prior Reuse</td>
<td>Citation count</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data citation</td>
<td></td>
</tr>
<tr>
<td>Advice on Reuse</td>
<td>Classes / education</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Documentation on how to reuse</td>
<td></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Transparency and clarity around data access processes (restricted)</td>
<td></td>
</tr>
</tbody>
</table>
Example features of a “reuser-oriented” data repository
Methodology

“These details are important, because data collection processes are highly individualized across quantitative and qualitative disciplines (Carlson and Anderson, 2007).”
"[Reusers] examine different factors such as the topic, the level of analysis, and the type of data to judge whether data is suitable or not to their purpose." (Curty, 2016)
“Data formats can indicate the degree to which the data are accessible, given the software data reusers have available or are skilled at using (Yoon, 2016)....”
“Data producer information, such as the names of the data creators, their institutional affiliations and where they were educated....” (Faniel, Frank, & Yakel, 2019).
“a lot of the work I’ve been focusing on actually is trying to understand uncertainty .... Not so much in the primary data itself but in this associated chronology which is necessary to interpret the primary data.” (Faniel, Frank, & Yakel, 2019).
Code for Reproducibility

Persistent link for this dataset:
https://doi.org/10.13012/B2IDB-0170190_V3

Related Code
Lindsay Clark. "R functions for conversion of genetic data formats."
Lindsay Clark. "Functions to help make nice plots in R."

Related Dataset
"Genetic diversity of Miscanthus sinensis." Raw DNA sequencing data
"Genetic diversity of Miscanthus sacchariflorus." Raw DNA sequencing data
"Genetic diversity and hybridization of Japanese Miscanthus sinensis and Miscanthus sacchariflorus." Raw DNA sequencing data
"Genetic diversity of Miscanthus in eastern Russia." Raw DNA sequencing data

Related Article
"I'll typically go and download the article and see what were they doing with the data. Again, it gets you before you start your own analysis a sense of what's possible, what's not possible..." (Faniel, Frank, & Yakel, 2019).

<table>
<thead>
<tr>
<th>Title</th>
<th>Semichemical responsive olfactory sensory neurons are sexually dimorphic and plastic (p56-IP-Seq of whole olfactory mucosa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>Experiment type</td>
<td>Other</td>
</tr>
<tr>
<td>Summary</td>
<td>Understanding how genes and experiences work in concert to generate phenotypic variability will provide a better understanding of individuality. Here, we considered this in the context of the main olfactory epithelium, a chemosensory structure with over a thousand distinct cell types, in mice. We identified a subpopulation of at least three types of olfactory sensory neurons, defined by receptor expression, whose abundances were sexually dimorphic. This subpopulation of olfactory sensory neurons was over-represented in sex-separated female mice and responded robustly to the male-specific semichemical 2-sec-buty-1,5-dihydrothiazole and (methythio)methanethiol. Sex-combined housing led to a robust attenuation of the female over-representation. Testing of Bax/- mice revealed a Bax-dependence in generating the sexual dimorphism in sex-separated mice. Altogether, our results suggest a profound role of experience and activity in influencing homeostatic mechanisms to generate a robust sexually dimorphic phenotype in the main olfactory epithelium.</td>
</tr>
<tr>
<td>Overall design</td>
<td>Phosphorylated S6 ribosomal subunit capture and associated mRNA profiling (p56-IP-Seq) on whole olfactory mucosa after odorant exposure.</td>
</tr>
</tbody>
</table>
“...all of the secondary data that I used is federally funded, is nationally representative, and is very widely used among the fields that I’m working in so I’ve never had to make like an independent judgment of, ‘Do I trust this survey or not?’...I just rely on the collective judgment to say that this a trustworthy survey.” (Faniel, Frank, & Yakel, 2019)
Links to Related Articles

New York City Taxi Trip Data (2010-2013)

Citation:
Donovan, Brian; Work, Dan (2016): New York City Taxi Trip Data (2010-2013). University of Illinois at Urbana-Champaign. https://doi.org/10.13012/J8PN93H8

Persistent link for this dataset:
https://doi.org/10.13012/J8PN93H8


Related Article

Related Code
https://github.com/UrbanTransport/parisresilience

Related Dataset
Donovan, Brian; Mori, Alec; Agrawal, Nimit; Meng, Yalan; Lee, Jong; Work, Daniel (2015): New York City Hourly Traffic Estimations (2010-2013). University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-4900670_V1

Related Thesis

Related Article

Related Article
“to know what’s restricted and what’s not restricted and then how to launch those procedures for getting that would be very helpful.” (Faniel, Frank, & Yakel, 2019).
“[Reusers] mentioned getting advice about working with data sets through workshops and courses....also expressed appreciation for documentation that included recommendations for working with the data (e.g. combining, weighting or linking data, etc.).” (Faniel, Frank, & Yakel, 2019).
Support Services

“The repository’s institutional affiliation, extent to which it was well known and widely used, collection size, staff size, expertise, and quality of curation and digitization work were used to assess its reputation and decide whether the data were credible and could be trusted.” (Faniel, Frank, & Yakel, 2019).
Versioning

"I know it’s Version 4, but I don’t know why [...] they [the data producers] could go, “Look,” it’s like a log saying, “We noticed that [...] there was a problem with blah-blah-blah [...] And we fixed this in variable blah-blah-blah. And this is Version 4”" (Faniel, Frank, & Yakel, 2019).
“Descriptions of how data, specimens, or artifacts were curated or digitized, including the people, functions, and/or services associated with these activities.” (Faniel, Frank, & Yakel, 2019).
<table>
<thead>
<tr>
<th>Data Repository Features Supporting Reuse Context</th>
<th>Data Repository Features Supporting Reuse Context</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Categories</strong></td>
<td><strong>Categories</strong></td>
</tr>
<tr>
<td><strong>Data Collection</strong></td>
<td><strong>Provenance</strong></td>
</tr>
</tbody>
</table>
| Methodology.  
Ephemeral documentation  
**Data type**  
Granularity or resolution | By source  
By traceability  
**Versions**  
Linkages  
File naming convention |
| **Specimen and Artifact**                     | **Repository information**                    |
| Taxonomies  
Provenance/source  
Unique IDs  
Access info  
Age estimation | Support services  
**Staffing levels** |
| **Data Producer**                             | **Curation and Digitization**                 |
| ORCID  
Other Data Producer | Documentation about the  
digitalization and curation  
process. |
| **Data Analysis**                             | **Prior Reuse**                               |
| Code book  
Documentation of data analysis  
**Reproducibility** | **Citation count**  
Data citation |
| **Missing Data**                              | **Advice on Reuse**                           |
| Documentation Procedure | **Classes / education**  
Documentation on how to reuse |
| **Research Objectives**                       | **Terms of Use**                              |
| Linkages research  
Linkages funder  
Abstract  
Data citations | Transparency and clarity  
around data access processes  
(restricted) |
Key Takeaways

➢ Some context types needed by reusers may not be easily accessible (documentation for missing data, repository curation services, etc.)

➢ Many data repositories are developed/tested for the depositor. But as a repository matures, we must shift our focus toward the reuser.

➢ New persistent identifier standards and linking technologies open up options for enhanced context in DRs. This evolving technology requires regular innovation and investment on the part of sponsoring institutions.

➢ This is preliminary research and we plan to do more interviews with disciplinary or generalist repositories.
Acknowledgements

The “Dissemination Information Packages for Information Reuse” (DIPIR) Project made possible in part by a National Leadership Grant from the Institute of Museum and Library Services, LG-06-10-0140-10, with additional support from OCLC and University of Michigan. https://www.oclc.org/research/areas/user-studies/dipir.html

The Data Curation Network is funded by the Alfred P Sloan Foundation and 12 members institutions. http://datacurationnetwork.org


Questions?

Thanks!

Contact us!

Ixchel M. Faniel, fanie@oclc.org
Lisa Johnston, ljohnsto@umn.edu
Katie Wissel, katie.wissel@nyu.edu