The Care and Feeding of Scientific Data

Mercè Crosas @mercecrosas
Director of Data Science, IQSS, Harvard University
On Data Sharing: 
What researchers want and what researchers do
Online survey with 1315 respondents across disciplines (9% response rate, mostly members of DataONE):

- Use others' datasets if their data were easily accessible: 84%
- Willing to share data across a broad group of researchers: 81%
- It is appropriate to create new datasets from the shared data: 76%

But 36% agreed that their own data are easy to access

But 46% shared some data, and only 6% shared all data

Ten-year study with 22 random participants from the Center for Embedded Network Sensing (CENS):

“Data sharing tends to occur only through interpersonal exchanges.”

“10 of the 22 participants were unaware of repositories that would accept data from their type of research.”

“14 participants said that they use data they themselves did not generate.”


Data sharing is mostly demand-driven
Survey sent to ~ 350 researchers at the Harvard-Smithsonian Center for Astrophysics; 175 respondents:

Have you ever used DATA you learned about from reading a Journal article?

- manually entered data from a table in a paper: 122 (19%)
- manually extracted data point values from a graph: 96 (15%)
- downloaded e-table of ASCII data provided by Journal: 113 (18%)
- contacted author to ask for data & got what I needed: 102 (16%)
- contacted author to ask for data & did NOT get what I needed: 46 (7%)
- used online archive where data were available: 150 (24%)


Data are accessed in various ways for reuse
When it comes to sharing DATA you’ve created, collected or curated, you have?

- emailed data to a colleague upon request. 151 (28%)
- put data at an ftp-style site for a colleague to retrieve. 104 (19%)
- put data at a personal web site 93 (17%)
- put data at a project-based web site 64 (12%)
- put data at an organized institutional archive 43 (8%)
- not shared my data, because I think it will endanger my career. 6 (1%)
- not shared my data due to large file sizes 21 (4%)
- not shared my data because I don't know how. 7 (1%)
- not shared my data because it takes too much effort. 22 (4%)
- not shared my data because I don't think anyone will want it. 23 (4%)
- Other 13 (2%)

I’ll share my data when you ask me
Links to data from 4 astronomy journals over 10 yrs
After 10 yrs since publication, >70% broken links
We can do better
10 Simple Rules

1. Love your data, and let others love it too
2. Share your data online, with a permanent identifier
3. Conduct science with data reuse in mind
4. Publish workflow as context
5. Link your data to your publications as early as possible
6. Publish your code
7. Say how you want to get credit for your data
8. Foster and use data repositories
9. Reward colleagues who share their data properly
10. Help establish data science and data scientist as vital

Goodman, Pepe, Blocker, Borgman, Cranmer, Crosas, Di Stefano, Gil, Groth, Hogg, Kashyap, Hedstrom, Mahabal, Siemiginowska, Slavkovic (2014), 10 Simple Rules for the Care and Feeding of Scientific Data, PLOS Computational Biology
A two-pronged approach to motivate cultural and policy change:

- Engage in policy debate, participate in community initiatives, and write papers like the “10 Simple Rules”
- Provide technical solutions to facilitate data sharing, reusability and interoperability
About Us

Data Science at IQSS combines expertise in software engineering, statistical innovation and data curation. Meet our team.

Current Efforts

Reproducible and Reusable Science
Connecting research results to the underlying data and analysis is central to the validation and extensibility of scientific discoveries. Our tools encourage open data and methodological transparency, when possible, and promote and enable data citation.

Computationally Assisted Exploration
We build analytical tools, such as Consilience and TwoRavens, that assist a researcher to understand and discover new insights from their data by connecting their own knowledge, expertise and judgement with the vast array of quantitative methods available in computational analysis.

Interdisciplinary Quantitative Scientific Scope
While social science research informs many of our software tools, we also engage with other fields such as physics, biology, and the social sciences to explore the impact and future of research in the digital age.

Software Projects

Zelig
Everyone's Statistical Software
Zelig: Everyone's Statistical Software is an interface, that allows a large body of different statistical models in the R statistical language to be implemented and interpreted in a common framework and interface.

Dataverse
Project
For almost a decade, Dataverse has been at the forefront of data publication, citation and preservation. We continue to innovate and improve Dataverse with the help of our users and partners.

Data Science Blog

Data Science Team Presenting at JavaOne!
Dataset Templates & Reset Password
Dataverse 4.0 Updates: More Metadata and SPSS File Handling
More ▸

TheData on Twitter

namsserc @thedatataorg Fantastic to hear Liz Quigley talk about usability today at Simmons. On open licensing, “That’s just how we roll.” Yes!
16 hours 23 min ago.


thedatataorg Data Science Team Presenting at JavaOne! We're excited to share our latest developments and insights in data science. Join us at this year's JavaOne conference!
### IQSS Data Science Team members

**Mercè Crosas**, Director of Data Science

<table>
<thead>
<tr>
<th>Statistics and Analytics</th>
<th>Software Development</th>
<th>Data Curation and Archivists</th>
</tr>
</thead>
</table>
| **James Honaker**, senior research scientist  
(Zelig, TwoRavens, RBuild) | **Gustavo Durand**, development manager  
(Dataverse) | **Sonia Barbosa**, archive and curation manager |
| **Christine Choirat**, research scientist  
(Zelig) | **Leonid Andreev**, senior software developer  
(Dataverse) | **Eleni Castro**, research coordinator, metadata specialist |
| **Vito d’Orazio**, postdoc  
(Zelig, TwoRavens) | **Phil Durbin**, software developer  
(Dataverse) | **Dwayne Liburd**, archivist |
| **Muhammed Idris**, predoc  
(Zelig, TwoRavens) | **Steve Kraffmiller**, software developer  
(Dataverse) | **Usability and User Experience** |
| **Quality Assurance and Technical Support** | **Michael Bar-Sinaï**, architect and senior software developer  
(Dataverse, DataTags, TwoRavens) | **Elizabeth Quigley**, usability specialist |
| **Kevin Condon**, QA and support lead  
(Dataverse, DataTags, TwoRavens) | **Raman Prasad**, BARI software developer  
(Dataverse, WorldMap) | **Michael Heppler**, UI designer & developer |
| **Elda Sotiri**, QA, technical support  
(Consilience, Dataverse) | **Robert Treacy**, architect and senior software developer  
(Consilience) | |
| | **Ellen Kroffmiller**, senior software developer  
(Consilience) | |
Dataverse: A bridge between traditional archives and posting data in your website

**Traditional data archives**
- Professional curation
- Full preservation
- Infrastructure to curate and preserve data
- Persistence guaranteed by hosting institution
- Tools to facilitate curation and preservation

**Posting data on the web**
- No curation or preservation guaranteed
- control and credit for data author
Dataverse Community

Federated Dataverses around the world with **persistence guaranteed by:**

- Dataverse.org coming at the end of 2014
- Dataverse advisory team and community groups:
  - API: common repository deposit API; search and data API
  - Metadata: standards per domain; automate extraction
  - Storage: multiple storages; integrate with iRODS
  - Preservation: integrate with archival and preservation tools
  - Authentication: multiple identity providers
  - Internationalization: chinese, spanish
Upcoming software improvements and new features
Dataverse 4.0, end of 2014

<table>
<thead>
<tr>
<th>Dataverse 4.0, end of 2014</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GBT Perseus HI Datacube</strong></td>
</tr>
<tr>
<td><strong>Aug 13, 2014</strong>  COMPLETE Dataverse</td>
</tr>
<tr>
<td><strong>COMPLETE team, 2014, &quot;GBT Perseus HI Datacube&quot;, <a href="http://dx.doi.org/10.5072/FK2/20">http://dx.doi.org/10.5072/FK2/20</a>, Harvard Dataverse, V1</strong></td>
</tr>
<tr>
<td>21 cm HI maps obtained at the 100 m NRAO Green Bank Telescope. The main component of HI emission toward the line of sight of Perseus is centered around 4 to 8 km s⁻¹, with the velocity of peak emission...</td>
</tr>
<tr>
<td>Subject: <strong>Astronomy and Astrophysics</strong></td>
</tr>
</tbody>
</table>

| Replication Data for: CSO/Bolocam 1.1-mm continuum in Ophiuchus |
| Aug 13, 2014  COMPLETE Dataverse |
| Young, Kaisa; Enoch, Melissa L.; Evans II, Neal J.; Glenn, Jason, 2014, "Replication Data for: CSO/Bolocam 1.1-mm continuum in Ophiuchus", http://dx.doi.org/10.5072/FK2/22, Harvard Dataverse, V1 |
| Data were taken May-June 2003 and 2004. Flux units are in mJy per 31 arcsecond beam. |
| Subject: **Astronomy and Astrophysics** |

| Replication Data for: CSO/Bolocam 1.1-mm continuum in Serpens |
| Aug 13, 2014  COMPLETE Dataverse |
| Enoch, Melissa L.; Glenn, Jason; Evans II, Neal J.; Sargent, Anelia I., 2014, "Replication Data for: CSO/Bolocam 1.1-mm continuum in Serpens", http://dx.doi.org/10.5072/FK2/23, Harvard Dataverse, V1 |
| Data were taken May-June 2003 and 2005. Flux units are in mJy per 31 arcsecond beam. |
| Subject: **Astronomy and Astrophysics** |

**More...**
A Dataset may contain any type of files, including code
Extensive Metadata, with data reuse in mind

- Descriptive metadata
  - **Citation Metadata** for all (compliant with DataCite)
  - **Domain metadata** blocks:
    - Social Sciences (compliant with DDI)
    - Biomedical (compliant with ISA-Tab)
    - Astronomy (compliant with VO)
    - Custom

- File Level metadata
  - **Automated extraction** of variables/columns metadata from R data, Stata, SPSS, Excel, CSV, and header metadata from FITS
Automated Data Processing

- RData
- Stata
- SPSS
- Excel
- CSV

Processing
- Extract metadata
- Re-format
- Calculate Numerical Fingerprint

Metadata File (XML, JSON) with column information

Data Table in Preservation Format
Data Exploration and Analysis Tools

- Tabular data
  - TwoRavens: Statistical analysis

- Data with geo-references
  - WorldMap: Statistical analysis

- Survey data
  - Survey Tool: cross-tabulations and reports

- Data with time variable
  - Time-series Visualizations: explore time series data
Open Licenses and Terms of Use

Multiple levels of access and reuse:

- Open License (CC0), with an understanding that scientific communication is based on attribution
- Custom Terms of Use
- Metadata open and files restricted: access may be granted upon request
On going collaborations
Automated Data Publishing

Integration of publishing systems with data repositories via API

Towards a common API across repositories and publishing systems

Journal Publishing System

Journal Dataverse
• Connect data to data (by analyzing metadata and usage)
• Connect data to users (via ORCID)
Data Citation and Provenance

- Incorporate provenance in data citation:
  - As metadata
  - DOI to provenance object

- Tracking multiple transformations:
  - disclosed provenance (e.g., explicit SQL query)
  - observed provenance (e.g., functions executed in R)
Sharing Sensitive Data
Thank you

@mercecrosas