Building and Supporting Data Repositories

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science (IQSS)
Harvard University
@mercecrosas mercecrosas.com

Qualitative Repositories Workshop, SESYNC, February 27, 2017
The importance of being FAIR
The importance of being cited
The importance of being connected
Data should be Findable, Accessible, Interoperable, Reusable (FAIR) by machines

To be Findable:
- global, persistent ID
- registered, indexed

To be Accessible:
- open, standard protocol
- open metadata

To be Interoperable:
- references to other metadata
- FAIR vocabularies

To be Reusable:
- standard, rich metadata
- clear data licenses
- provenance

“FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.”

“Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.”
The importance of being FAIR
The importance of being cited
The importance of being connected
Today's Bibliographies and CVs


Altman, M., Crossas, M. 2013. The Evolution of Data Citation: From Principles to Implementation, IASSIST Quarterly, p. 62

Future Bibliographies and CVs

data sets


Repositories should implement data citation maximizing discovery and access

Required:
• persistent ID/url resolves to dataset landing page

Recommended:
• landing page includes human- and machine-readable metadata

Optional:
• content negotiation for more accessible metadata

Fenner et al, 2016, “A Data Citation Roadmap for Scholarly Repositories” BioArxiv (preprint)
Synthesis Group, 2014, Joint Declaration of Data Citation Principles (Force11)
The importance of being FAIR
The importance of being cited
The importance of being connected
To build incentives and impact, all parties need to be on board.

- **Publishers & Journals**: connect articles to data
- **Bibliographic repositories**: count data citations
- **Discovery Indexes & Registries**: find data
- **Data repositories**: open data
- **Researchers**: deposit data, get credit
- **Funders**: support data
- **Institutions**:
Dataverse

Also for Qualitative Data
We built Dataverse to incentivize data sharing, with good data management in mind

- An open-source platform to share and archive data
- Developed at Harvard’s Institute for Quantitative Social Science (IQSS) since 2006
- Gives credit and control to data authors & producers
- Implements FAIR Principles and Data Citation roadmap*
- Builds a community to:
  - define new standards and best practices
  - foster new research in data sharing and reproducibility
- Has brought data publishing into the hands of data authors
Dataverse is now a widely used repository platform

22 installations around the world
70,000 datasets in Harvard Dataverse repository
Used by researchers from > 500 institutions
http://dataverse.org
Basic Citation metadata required, rich metadata recommended.
**Terms of Use & Licenses:**

Open data encouraged, restrictions optional; **DataTags** to enable sharing sensitive data (coming soon).

**Description**

Bike collisions in Boston (2009-2012) were organized and compiled by the Boston Cyclists Union. Visitors are encouraged to report, as per the provided reports, as part of the dataset.

**Terms of Use**

Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse.

**Guestbook**

The following guestbook will prompt a user to provide additional information when downloading a file.

**Boston Area Research Initiative**
Datasets and file versions, with provenance (coming soon)
Thanks!
Learn more at http://dataverse.org

@mercecrosas  mercecrosas.com