Integrating Data Citation with Provenance

Margo Seltzer, Mercè Crosas, Gary King

HARVARD School of Engineering and Applied Sciences

IQSS The Institute for Quantitative Social Science

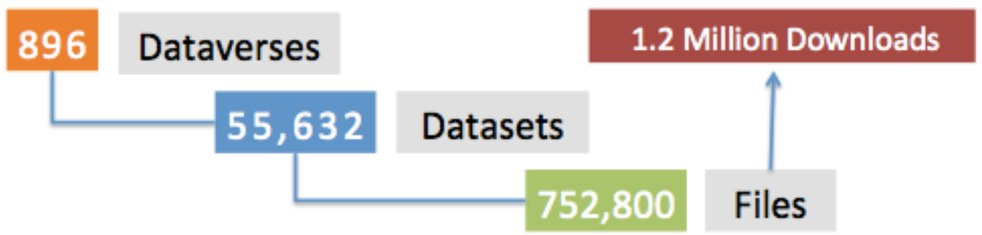
What is Dataverse?

Software framework for publishing, citing and preserving research data
(open source on github for others to install)

Provides incentives for researchers to share:

- Recognition & credit via **data citations**
- Control over data & branding
- Fulfill journal data availability and funder data management requirements

Harvard Dataverse (**open to all;** repository instance at Harvard) currently has:



Data Citation with Dataverse

Integrates with DataCite service for minting persistent identifiers (DOIs) and registering dataset citation metadata

Authors, Year, Dataset Title, DOI, Data Repository, UNF, version

Attribution to data authors and distributors

Fingerprint (UNF) to verify dataset, and version to specify what data are being referenced. UNF does **not** depend on the data format.

What we propose

- *Describe Provenance with **code** used to transform or merge datasets*
- *Represent **Provenance as a graph**, connecting the code with the input and output datasets (using DOIs)*
- *Include Provenance graph in Dataset **metadata***

Anticipated Collaborations:

- *With DataCite, enhance their metadata schema to support provenance graph*
- *With USENIX, integrate their open access repository with the citation + provenance work from this project*

Provenance examples with R Code