Dataverse
Sharing Research Data; Building Social Science

Gary King

Institute for Quantitative Social Science
Harvard University

Talk at the American Political Science Association meetings, August 2014
1. **Long term preservation:**
   - Following *international standards* of preservation formatting, documentation, attribution, code, metadata
   - Care by *professional archivists* (hint: that’s probably not you!)
   - A *trusted data repository*, backed by endowment or institution
   - An active, supportive *community*
     - Assist, not replace, existing archives
     - Open source development, user groups, installations worldwide

2. **Credit to the author (not only the archive):**
   - *Formal data citation*: including persistent identifier (DOI), universal numeric fingerprint (UNF), etc.
   - *Web visibility and Control*: a virtual archive with ownership by author (everyone gets 100% of the credit)
   - *Ever improving software*: sophisticated on-line analytics (Zelig), download in every format, high visibility for distribution, etc.
Options for Archiving Replication Data at Journals

**Option A: Do it yourself**
1. Publish data to authors’ own Dataverse; get citation
2. Cite data in paper; journal only needs to check for citation

**Option B: With journal verification**
1. Submit draft of replication data to Journal Dataverse
2. Journal reviews it, and approves it for release
3. Dataset is published with a formal data citation & back to the article.

**Option C: Full Automation**
1. Seamless Integration between journal submission system and Dataverse
2. Automatic Link created between article and data
3. Easy for journal and author, and less error-prone.
Journals can have their own Dataverses too.
Sloan funded project to integrate PKP’s Open Journal System (OJS) with Dataverse software via API

- Full integration thus far: ~400 journals (out of 5000 using OJS)
- Many more integrations in progress
An Integrated, Living Workflow

Submit

Authors to include data citations in Reference list

Integration with data repositories (repository API)

Prepare new submission

Dataset: permanent, with versioning, merging, improving

Journals can review data before article publication

Review

Data citation: finding & reusing data

Reuse & Extend

Data

Article

Submit

Authors to include data citations in Reference list

Integration with data repositories (repository API)

Prepare new submission

Dataset: permanent, with versioning, merging, improving

Journals can review data before article publication

Review

Data citation: finding & reusing data

Reuse & Extend

Data
• **Confidentiality, Privacy:**
  - DataTags: self-customizable data risk and sensitivity labels
  - Highly secure storage, with multiple security levels

• **New Data Challenges:**
  - Very large scale data sets
  - Streaming data (continuous versioning)
  - Tools for online fast data access, queries, and visualizations
  - Analysis methods for when data cannot be moved
  - Use (huge volume of) web analytics to improve dataverse and Zelig
For more information

TheData.org

GaryKing.org