

Harvard Dataverse Management and Processes

(TDR Webinar Series, 6/25/19)

SONIA BARBOSA

MANAGER OF DATA CURATION, HARVARD DATAVERSE



HARVARD
Dataverse

The
Dataverse[®]
Project



HARVARD
LIBRARY

Harvard Dataverse Processes

Type of Repository:

- Self-deposit, self-curated (managed, semi-managed)
- Support@dataverse.harvard.edu (Harvard Dataverse questions)
- Support@dataverse.org (Dataverse software questions)

Support Triage:

- IT Client Service at IQSS (level 1)
 - Curation Team (level 2)
 - Curation manager and Product Development manager (level 3)

Harvard Dataverse Processes

Harvard Dataverse Process

- Create your own account
- Manage your own team
- 2.5GB per file limitation
- 1 terabyte limitation (non-harvard affiliates)
- CC0 by default, user can change as needed
- All workflows supported
- Daily Review of deposited content (by Curation Management Team)

Harvard Dataverse Processes

Variables we check for:

- Spam
- Correct “category” of data
- “Test” datasets
- Sensitive data
- Published datasets without data (note in abstract when data are coming)
- Draft content (Catch issues before dataset published)
- Restricted content must have “Terms of Access” or enabled “request access”
- Backlog (>3 years : remove old empty dataverses, old unpublished datasets (with data, with no data, always requires a contact)

Harvard Dataverse Processes

Variables we check for:

- Is the “replication data for” feature being used correctly
- Is the dataset title complete
 - “Data for my study” is not an appropriate title
- How many data files are in the dataset and will it interfere with the “download” all functionality
- File Formats:
 - Can’t download
 - Passcode protected
 - Failed tabular ingest
 - WorldMap feature
- Are the default features we provide working correctly?
 - BUGS!

Harvard Dataverse Processes

For managed (large, good quality) collections we offer additional curation “advice:”

- Journals: have they selected the journal metadata block and facets?
- Customization, Subdataverses, Guestbooks, and Templates
- Many depositors don’t know how to use the “facets” set-up in creating a dataverse so they skip it altogether. The relationship between the “metadata block” and “facets” is often missed.

Other

- Multiple “dataset” deposits by one author?
 - The should create a “dataverse!”
- How should a dataset be organized?
 - Data! Documentation! Code! Supplementary Document as **one dataset!** and not multiple datasets!

Harvard Dataverse Processes

Other:

- Journals with only articles (not documentation as data, etc...) - automatically remove permissions and contact for a demo on how Dataverse is to be used, and what constitutes data. If no reply, dataverse and contents deleted
- Attention to default CC0 Terms on Harvard Dataverse
 - We point this out to managed, large collections. They can continue to use CC0 with additional details depending on their needs. For example, CC0 + “downloads should only take place on Harvard Dataverse.”

Harvard Dataverse Processes

How to handle Sensitive Data Protocols:

- No sensitive data until “DataTags (<http://datatags.org/>)”
- HIPAA regulations guidelines
- The title of the dataset is often a clue! E.g. “facebook,” “hospital,” “community survey,” etc...
 - General contact to the author to reiterate the sensitive data policy and asking them to confirm their data have been deidentified prior to upload.
- “Deaccessioning” used if sensitive data have been published and accessed by users
- “Deletion” possible if data have not been downloaded
- Any data that are near sensitive but cleaned, we recommend “guestbook” in case sensitive issues arise later, data user can be contacted (GDPR?)



Harvard Dataverse Processes

Outreach:

You control your data! You own your data!

- Citations, discoverability, increased use, permanent DOI (dataset, file-level), team access, easy data sharing, guestbook to track users and provide summary use stats for funding agency, etc...

We provide the data sharing “plumbing:”

- Tools, long-term preservation, support

The risks:

- Data loss
- 

Harvard Dataverse Processes

Some sustainable guidelines:

Template responses to common issues:

- Failed tabular ingest
- Sensitive data support contact
 - Have the depositor take responsibility for making sure data are clean
- Deaccessioning

Keep track of “active” bugs for your team and your users

Thank you!



HARVARD
Dataverse

The
Dataverse[®]
Project



HARVARD
LIBRARY