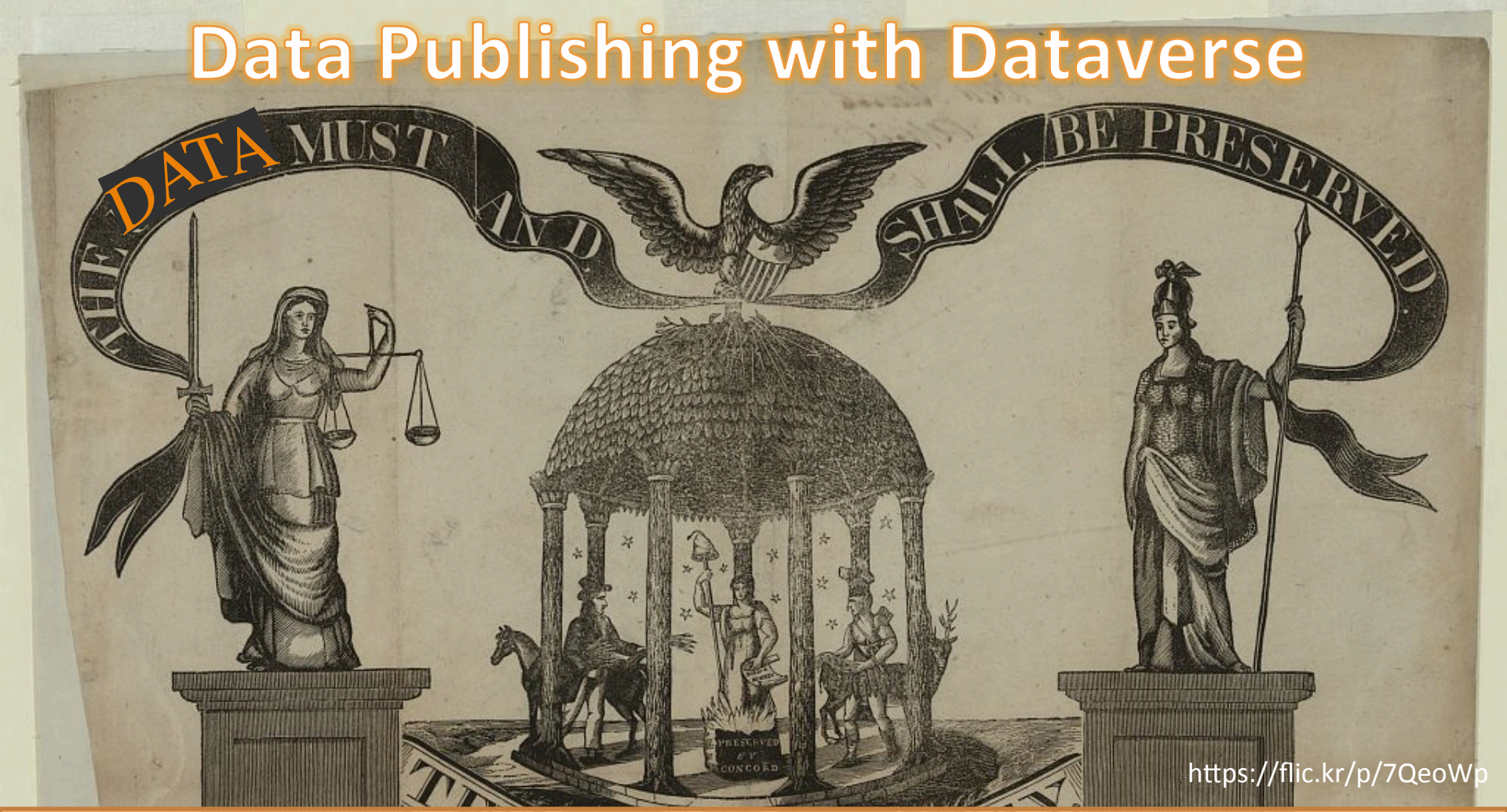
Data Publishing with Dataverse



https://flic.kr/p/7QeoWp

Eleni Castro, Research Coordinator
The Institute for Quantitative Social Science (IQSS), Harvard University
@thedataorg

IQSS
The Institute for Quantitative Social Science

NDSA NE Regional Meeting,
October 30, 2014 > Amherst, MA

The Dataverse Project

IQSS Data Science Team

The Institute for Quantiative Social Science


Researchers



Statistical Innovation


Data Science Applications and Tools

Data Curation & Stewardship

Information Scientists

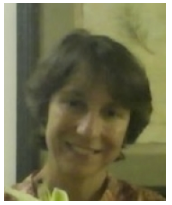
Tool Building & Computer Science


Software Engineers







Find out more: http://datascience.iq.harvard.edu

What is the Dataverse Project?

Software framework for Research Data (since 2006):

- Publishing
- Citing
- Analyzing
- Preserving

Open source on github.

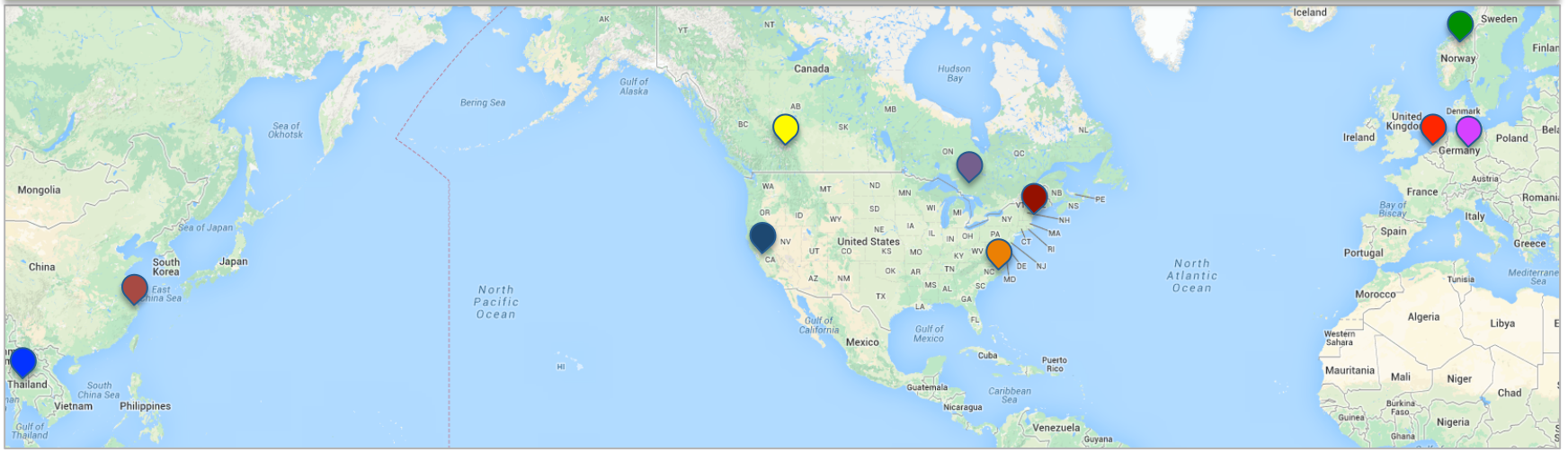
Provides incentives for researchers to share:

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

The Dataverse Project

Who Uses Dataverse ?

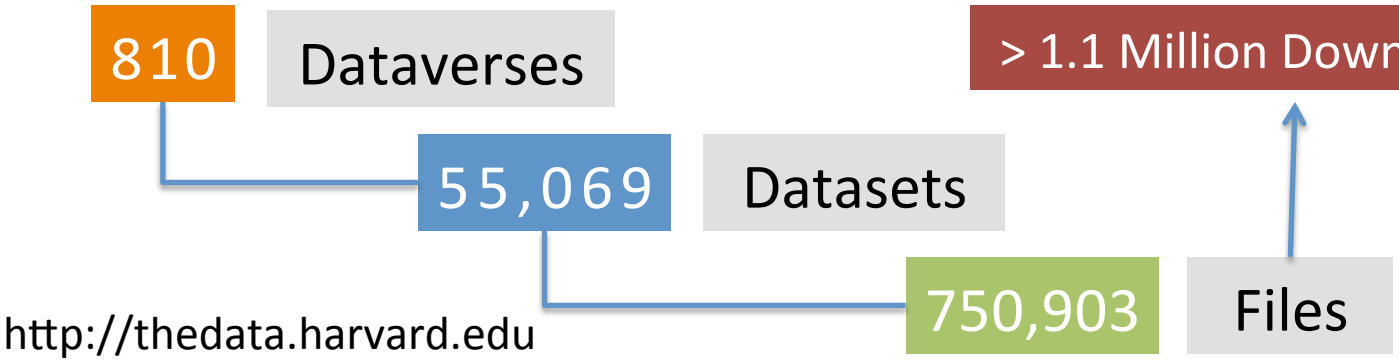
Worldwide Dataverse Installations



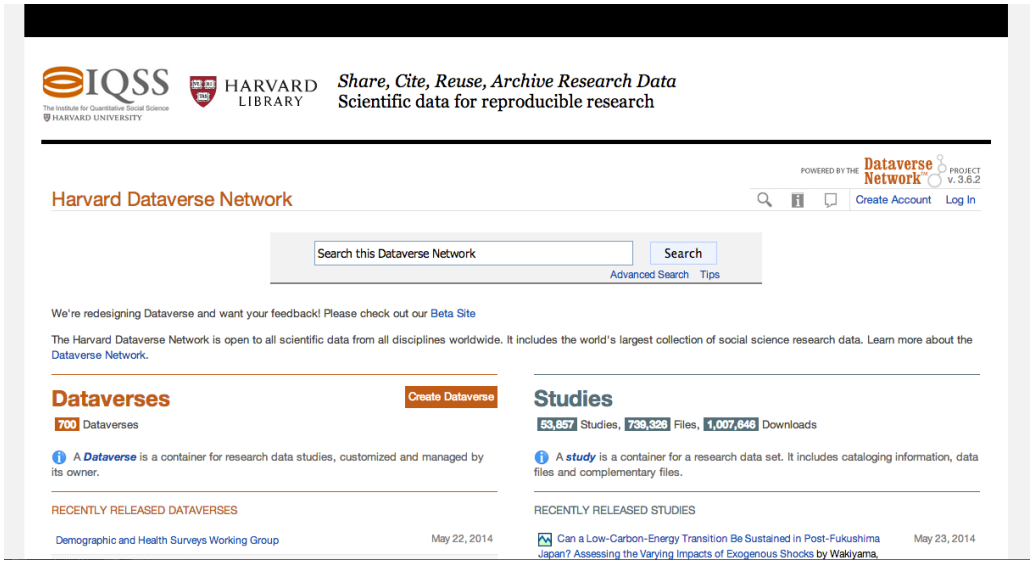
Institutions can setup/host their own Dataverse installation (UNC ODUM, Fudan Univ, Scholars Portal, DANS, etc) and within them can have dataverses for a variety of users (across all research domains): Researchers, Projects, Journals, etc.

Harvard Dataverse

Open to all researchers; general community repository instance:



http://thedata.harvard.edu



Harvard Dataverse Network >

POWERED BY THE Dataverse Network™ PROJECT v. 3.6.2

National Digital Stewardship Alliance Dataverse

Create Account Log In

Data archive for data collected by the national digital stewardship alliance.

Search Studies Search

Advanced Search Tips

National Digital Stewardship Alliance

Sort By: Global ID

Studies: 3 | Downloads: 20

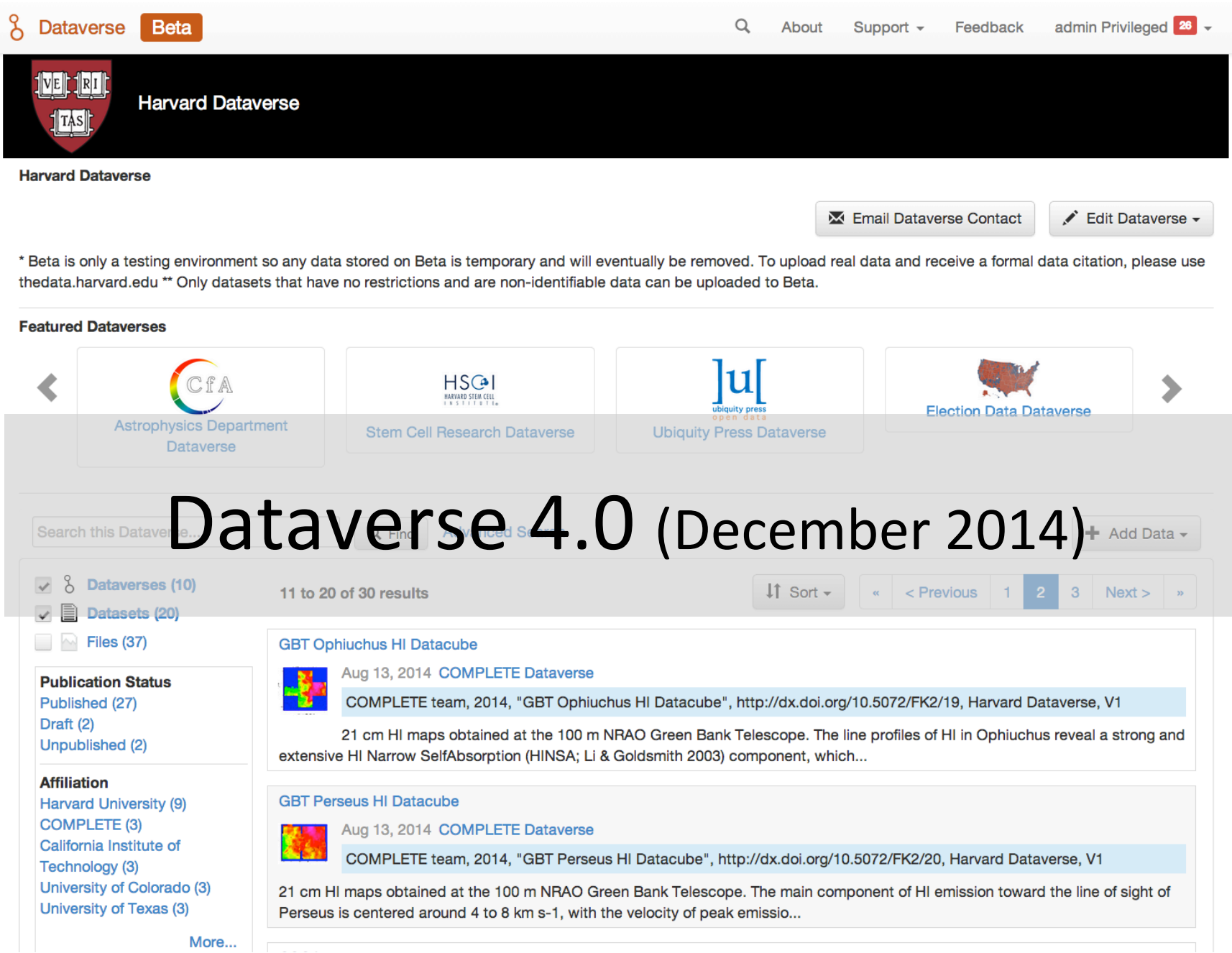
NDSA Web Archiving Surveys
by Content Working Group
Description: The NDSA conducted two surveys of organizations in the United States that are actively involved in, or planning to start, programs to archive content from the web. The goal of the survey was to better understand the landscape of web archiv…Continue [+]
Distribution Date:October 14, 2014
doi:10.7910/DVN/27593
8 downloads
Last Released: Oct 23, 2014

Replication data for: Staffing for Effective Digital Preservation: An NDSA Report
by National Digital Stewardship Alliance (NSDA). Standards and Practices Working Group.
Description: Businesses, cultural memory institutions, repositories, and government bodies seeking to preserve digital assets responsibly face significant staffing challenges. How many staff and what types of positions are required? What skills, educat…Continue [+]
Distribution Date:October 01, 2013
doi:10.7910/DVN/27562
Last Released: Oct 9, 2014

Data for NDSA Storage Report 2011
by Altman, Micah; Bailey, Jefferson; Cariani, Karen; Gallinger, Michelle; Owens, Trevor
Description: The NDSA storage survey aims to gather information on preservation storage systems. The respondents represent a diverse cross section of organizations working with preservation storage systems including university libraries, consortia, ins…Continue [+]
Distribution Date:2012
Related Publications:Micah Altman, Jefferson Bailey, Karen Cariani, Michelle Gallinger, Jane Mandelbaum, Trevor Owens, and the NDSA Infrastructure Working Group, 2013, Reflections on National Digital Stewardship Alliance Member Approaches to Preservation Storage Technologies: NDSA Storage Report, D-Lib Magazine. Forthcoming.
hdl:1902.1/19768
12 downloads
Last Released: Feb 10, 2013

NDSA.org

http://thedata.harvard.edu/dvn/dv/ndsa


Dataverse Beta
About
Support
Feedback
admin Privileged 26
Harvard Dataverse
Harvard Dataverse
Email Dataverse Contact
Edit Dataverse
* Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. To upload real data and receive a formal data citation, please use thedata.harvard.edu ** Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.
Featured Dataverses
Astrophysics Department Dataverse
Stem Cell Research Dataverse
Ubiquity Press Dataverse
Election Data Dataverse
Dataverse 4.0 (December 2014)
Add Data
Dataverses (10)
Datasets (20)
Files (37)
11 to 20 of 30 results
Sort
« < Previous 1 2 3 Next > »
Publication Status
Published (27)
Draft (2)
Unpublished (2)
Affiliation
Harvard University (9)
COMPLETE (3)
California Institute of Technology (3)
University of Colorado (3)
University of Texas (3)
More...
GBT Ophiuchus HI Datacube
Aug 13, 2014 COMPLETE Dataverse
COMPLETE team, 2014, "GBT Ophiuchus HI Datacube", http://dx.doi.org/10.5072/FK2/19, Harvard Dataverse, V1
21 cm HI maps obtained at the 100 m NRAO Green Bank Telescope. The line profiles of HI in Ophiuchus reveal a strong and extensive HI Narrow SelfAbsorption (HINSA; Li & Goldsmith 2003) component, which...
GBT Perseus HI Datacube
Aug 13, 2014 COMPLETE Dataverse
COMPLETE team, 2014, "GBT Perseus HI Datacube", http://dx.doi.org/10.5072/FK2/20, Harvard Dataverse, V1
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 emissio...

Dataverse Best Practices (1)

- Standard Metadata Schemas
  - DDI, FGDC & OAI DC
  - Coming in 4.0: DataCite 3.1, ISA-Tab (biomedical), and VO Resource (astronomy)→ Export to JSON & XML
- Formal Data Citation (Altman & King, 2007)
  - Endorse + comply w/ Joint Declaration of Data Citation Principles (FORCE11)
  - Versioning and File Fixity
- Persistent IDs: Handles & DOI (DataCite/EZID)

Dataverse Best Practices (2)

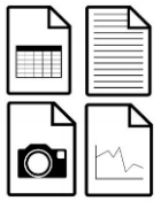
- Preservation format conversion for tabular data (extract column/variable metadata)
- File Fixity:
  - UNF (Altman, 2008) for tabular data
  - MD5 checksums for other files

Dataverse Best Practices (3)

- LOCKSS (replication of files)
  → Data-PASS: (ICPSR, ODUM, NARA, ROPER,…)

- OAI-PMH: Harvesting metadata (DC, DDI)
  – From other Dataverse installations
  – From other OAI-DC compliant repositories
- If necessary: Deaccession a Dataset

Rigorous Data Publishing Workflows



Upload

Draft Dataset

In 4.0: Major + Minor Versioning

Publish Version 1

Published Dataset v1

Authors, Title, Year, DOI, Repository, V1

Publish Version 1.1: small metadata change; citation doesn't change.

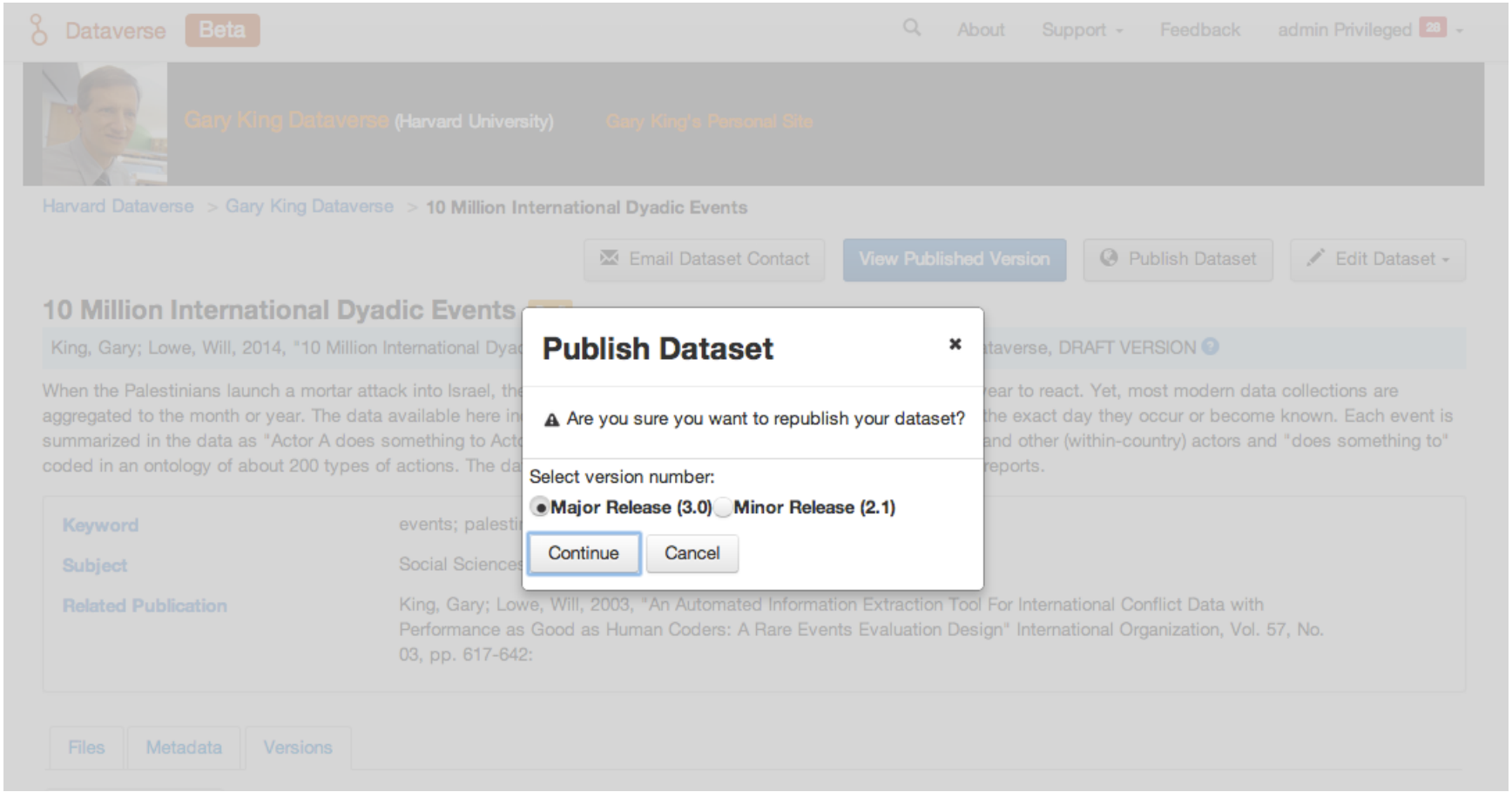
Published Dataset v1.1

Publish Version 2: File change (automatic); big metadata change (e.g., author, title).

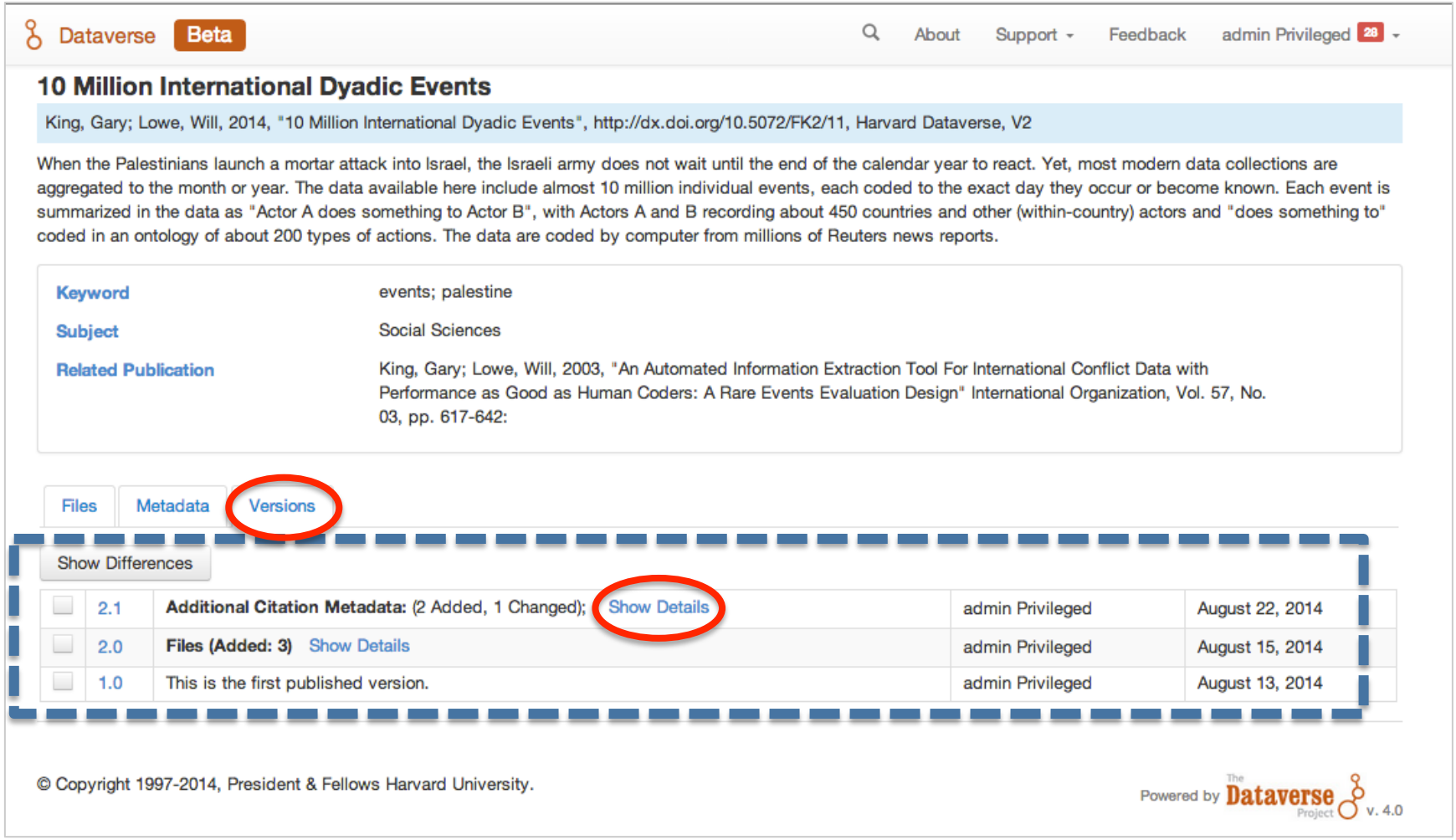
Published Dataset v2

Authors, Title, Year, DOI, Repository, UNF, V2

Dataset Versioning (1)



Dataset Versioning (2)



Deaccession Data in 4.0

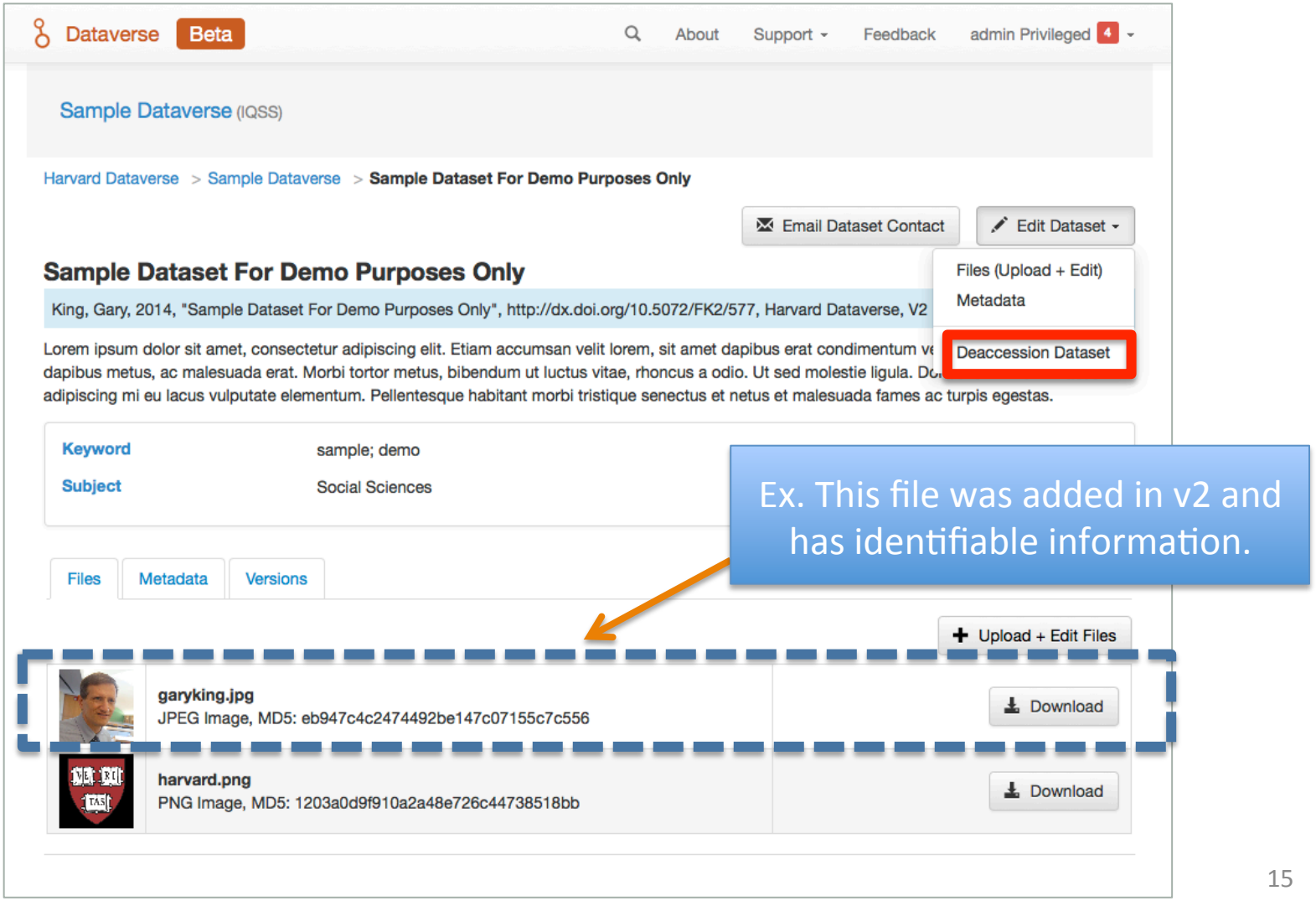
Before a Dataset is published the DOI is private (reserved). Only when published is it made public & searchable.

In accordance w/ **Data Citation Principle #6** Persistence: A Published Dataset **cannot** be deleted; only deaccessioned, with a reason.

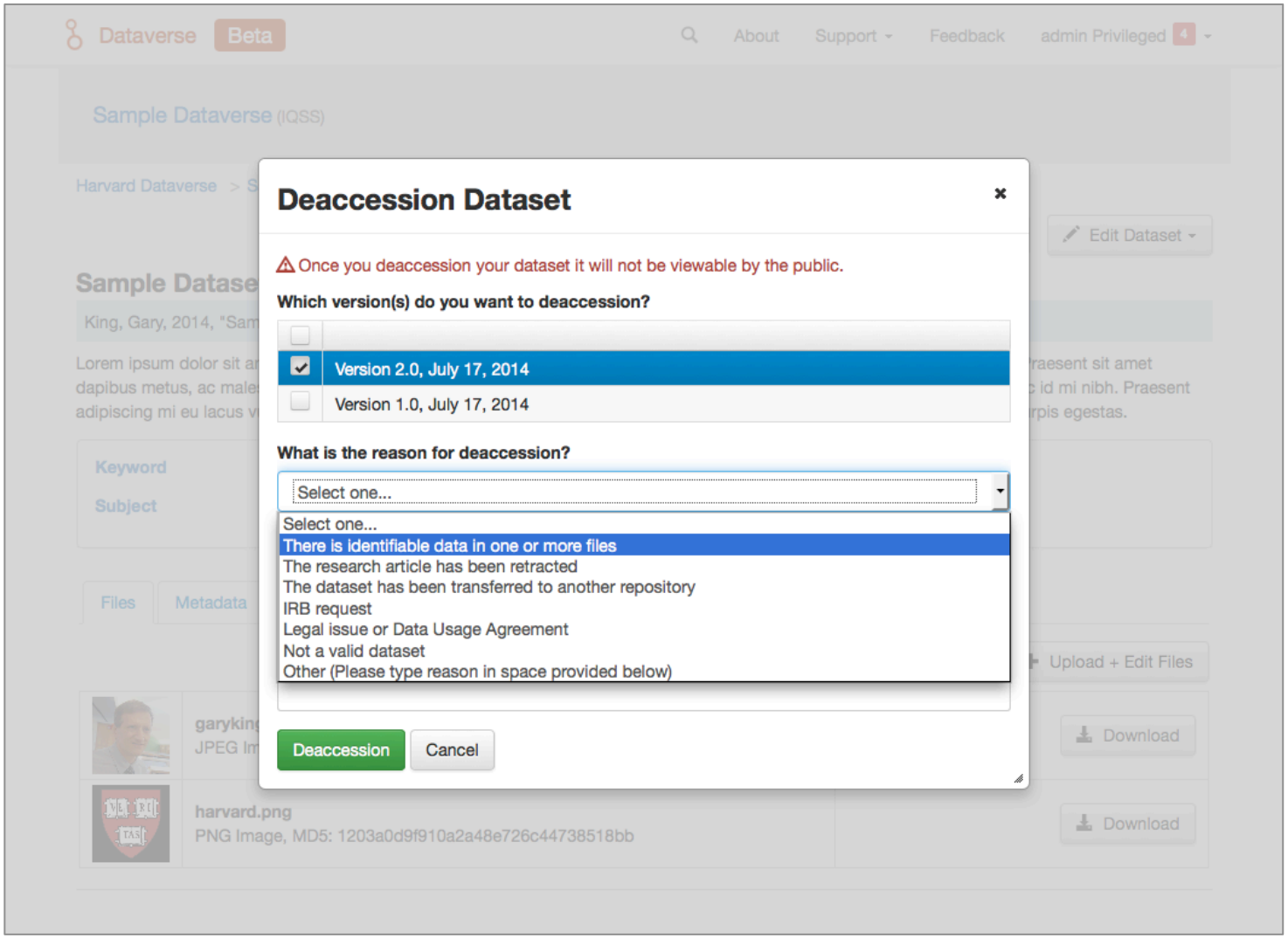
You can Deaccession (in 4.0):

1. **version**(s) of a Dataset, or
2. an **entire** Dataset.

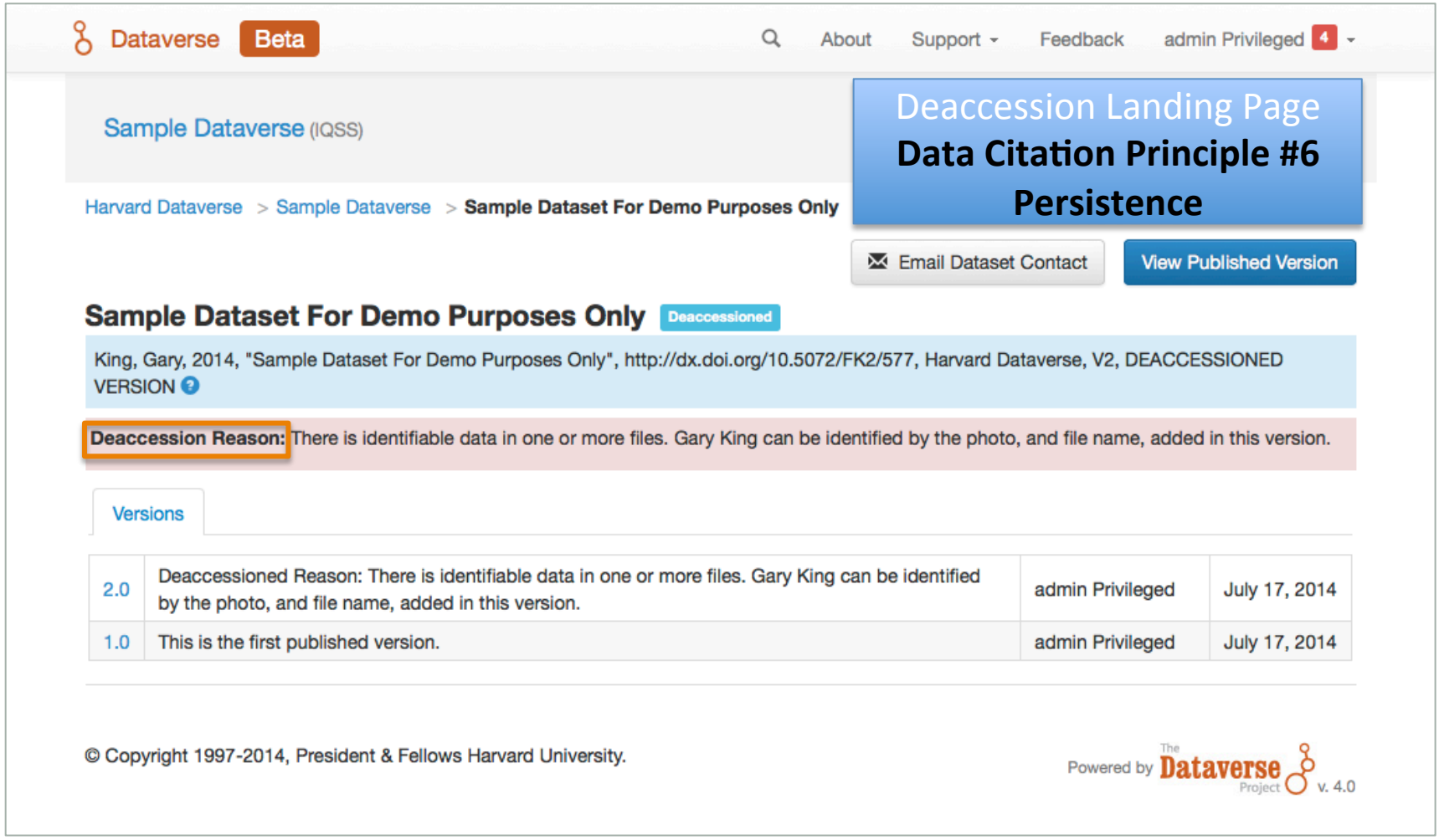
Deaccession A Version (1)



Deaccession A Version (2)



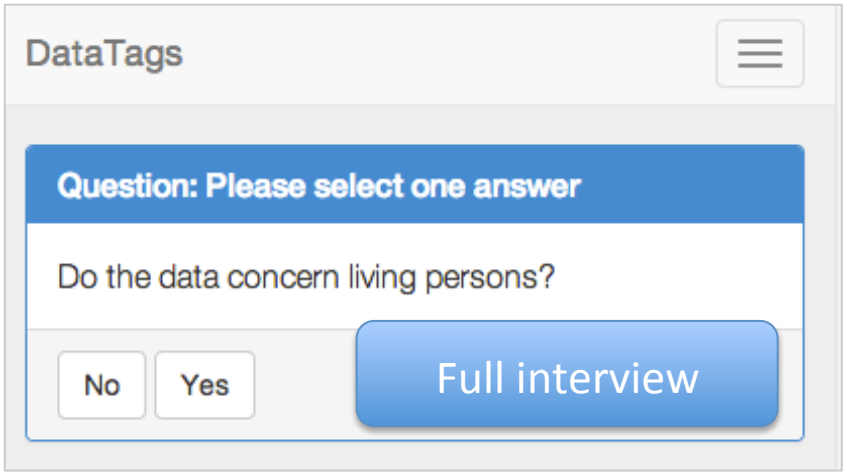
Deaccession A Version (3)

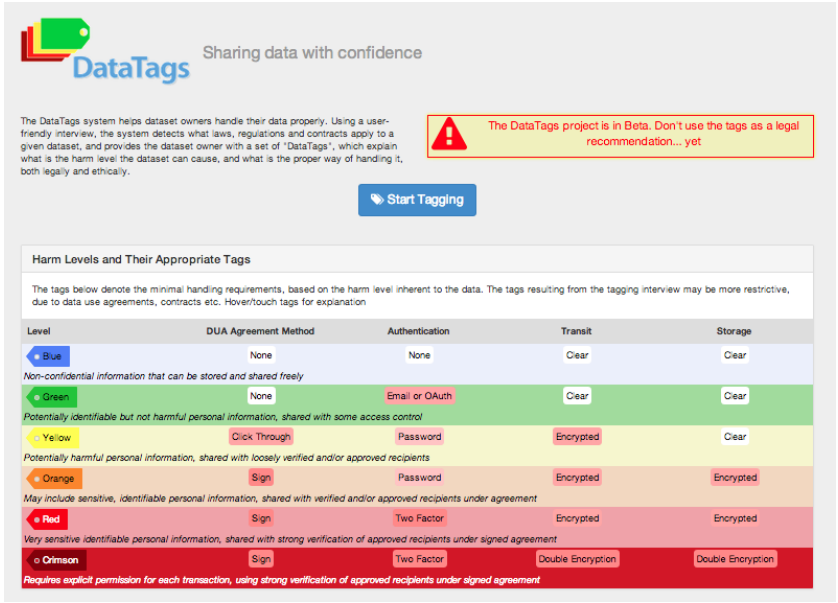


Data Publishing After 4.0 (2015)

Publishing Privacy Sensitive Data

- Secure Dataverse
- **DataTags (demo)** (based on Privacy Laws and DUAs)

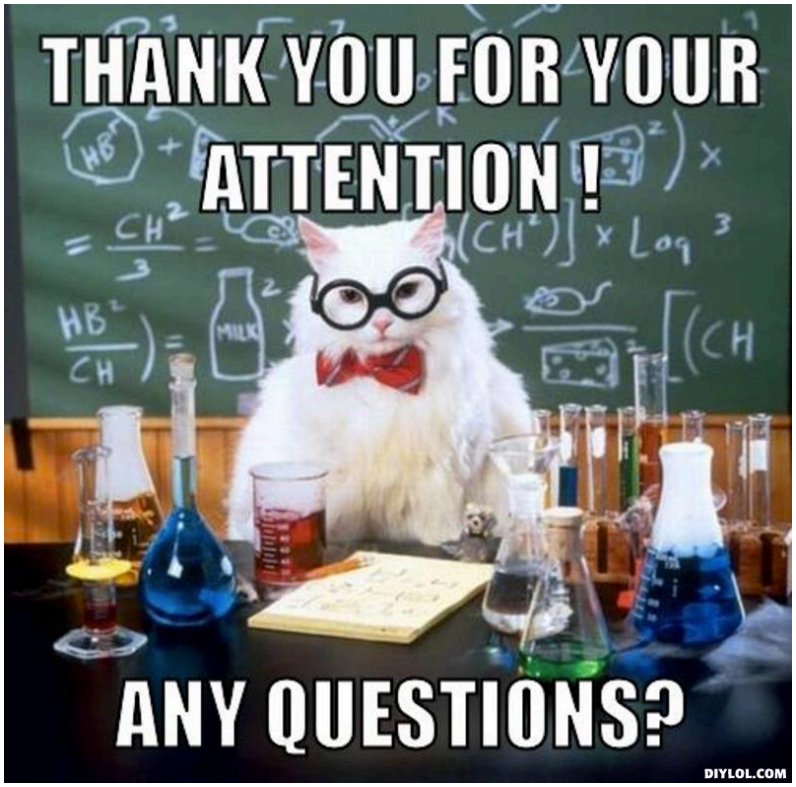




- Data Citation Provenance Registry (w/ SEAS funded by NSF)
- Author Disambiguation: ORCID Integration (API)
- Long-term preservation for more file formats (beyond tabular)
  - UDFR, Archivematica,…

Thank you!

Contact: ecastro@fas.harvard.edu



References

Altman M. A Fingerprint Method for Verification of Scientific Data. In A Fingerprint Method for Verification of Scientific Data. Springer-Verlag; 2008.

Altman M, King G. A Proposed Standard for the Scholarly Citation of Quantitative Data. D-Lib Magazine [Internet]. 2007;13(3/4).