The Dataverse Project

Introducing the Expanding Dataverse

Elizabeth Quigley
Usability Specialist
IQSS @ Harvard University

Introduction to Dataverse

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

Developed by the Institute for Quantitative Social Science at Harvard University.

Provides incentives for researchers to share:

- Recognition & credit via data citations
- Control over data & branding
- Fulfill Data Management Plan requirements
- Default CC0 Waiver for all uploaded datasets

Dataverse Milestones

1999-2006:
Virtual Data Center
(VDC)

→

2006:
Coding of Dataverse
begins (initial focus on
Social Sciences data)

→

2011:
Expanded to include
Astronomy &
Astrophysics data

→

Fall 2012:
OJS & Dataverse
API Integration

→

March 2012:
Dataverse 3.0 Released

→

April-May 2013:
First Usability Testing
of Dataverse

→

2013:
Expanded to include
Biomedical data

→

October 2013:
Dataverse 4.0
development begins

→

October 2013:
User centered design
process integrated

→

October 2013-April 2015
Dataverse 4.0
Development

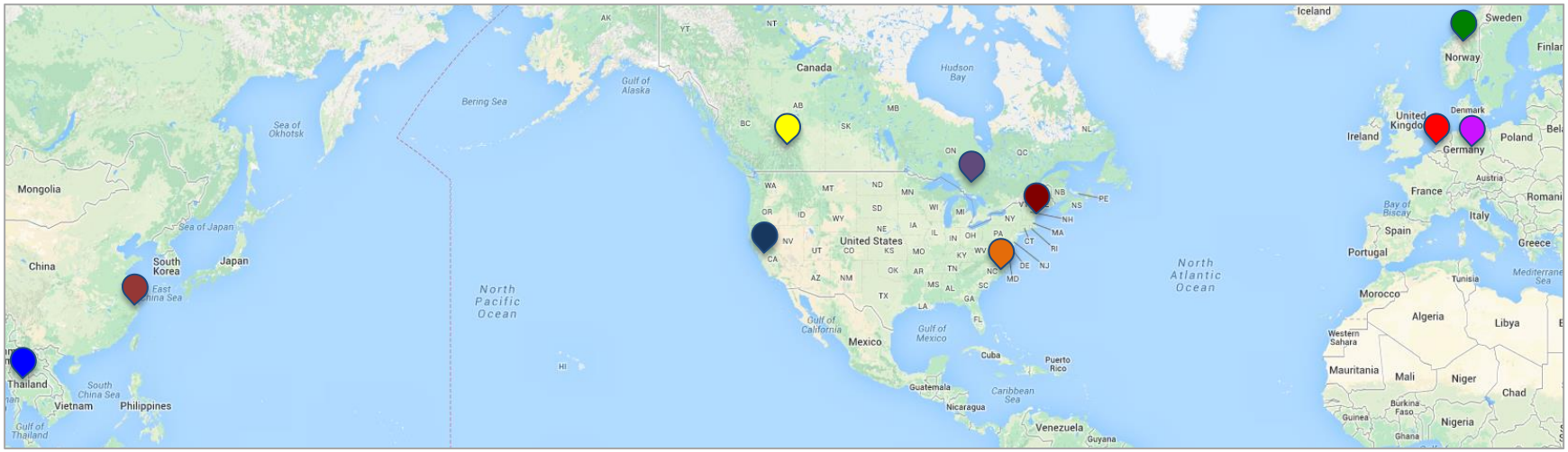
→

April 2015:
Dataverse 4.0 Launched
dataverse.harvard.edu

Who uses Dataverse?

- Researchers
- Librarians
- Data Archivists
- Journals
- Courses
- Institutions and Organizations

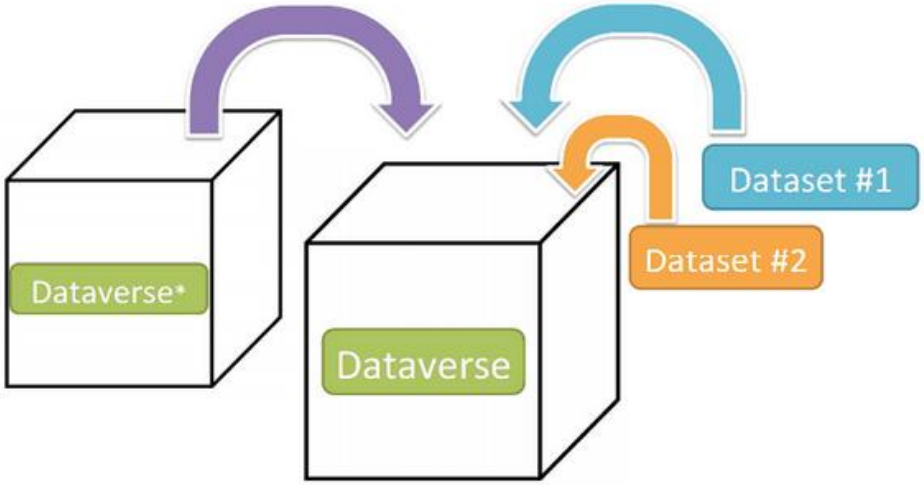
Dataverse Around the World



Institutions can setup/host their own Dataverse repository (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.

What is a Dataverse or Dataset?

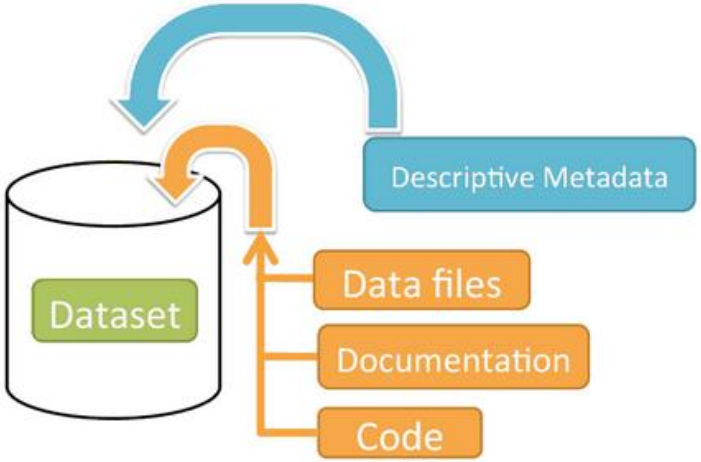
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

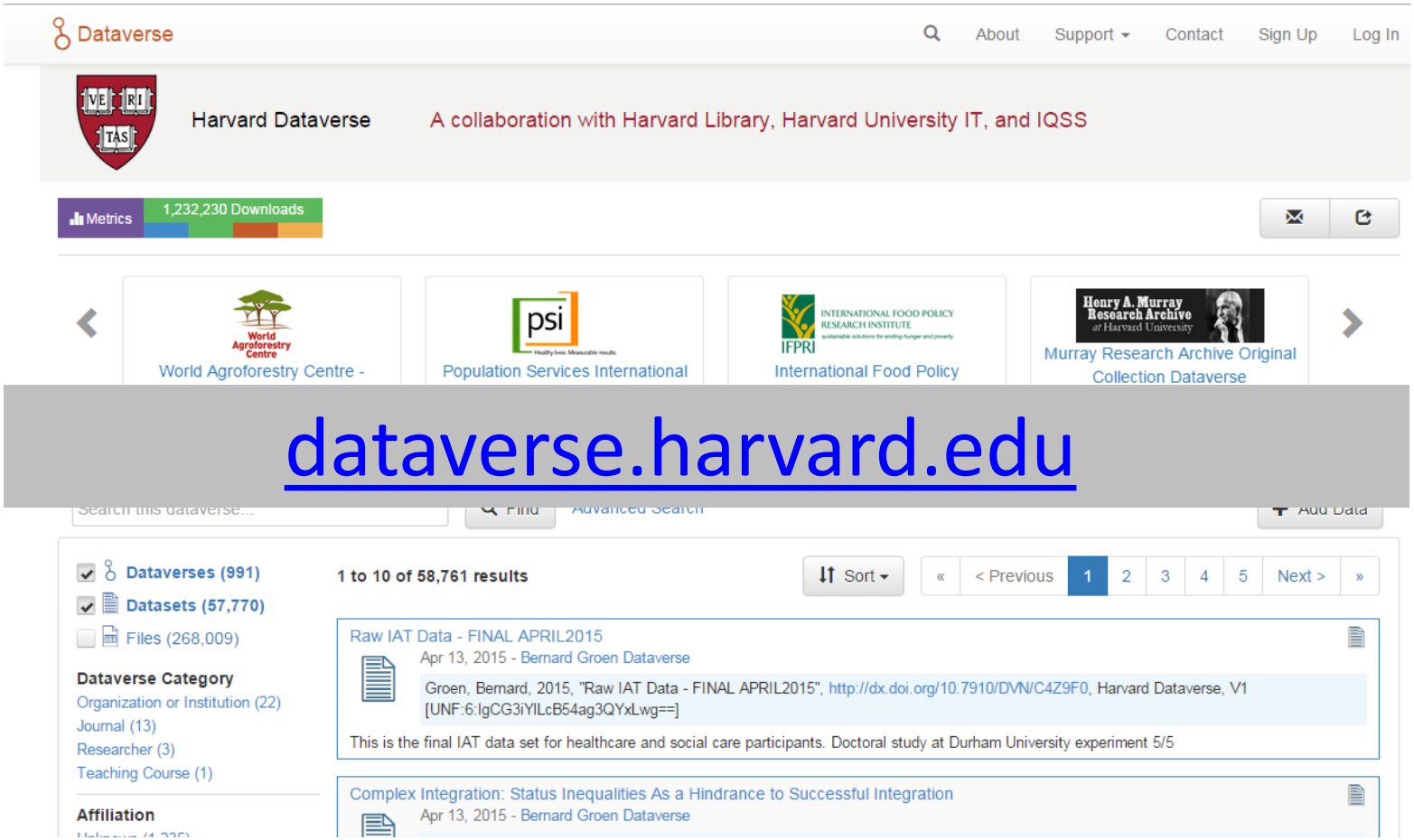
Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

Harvard Dataverse

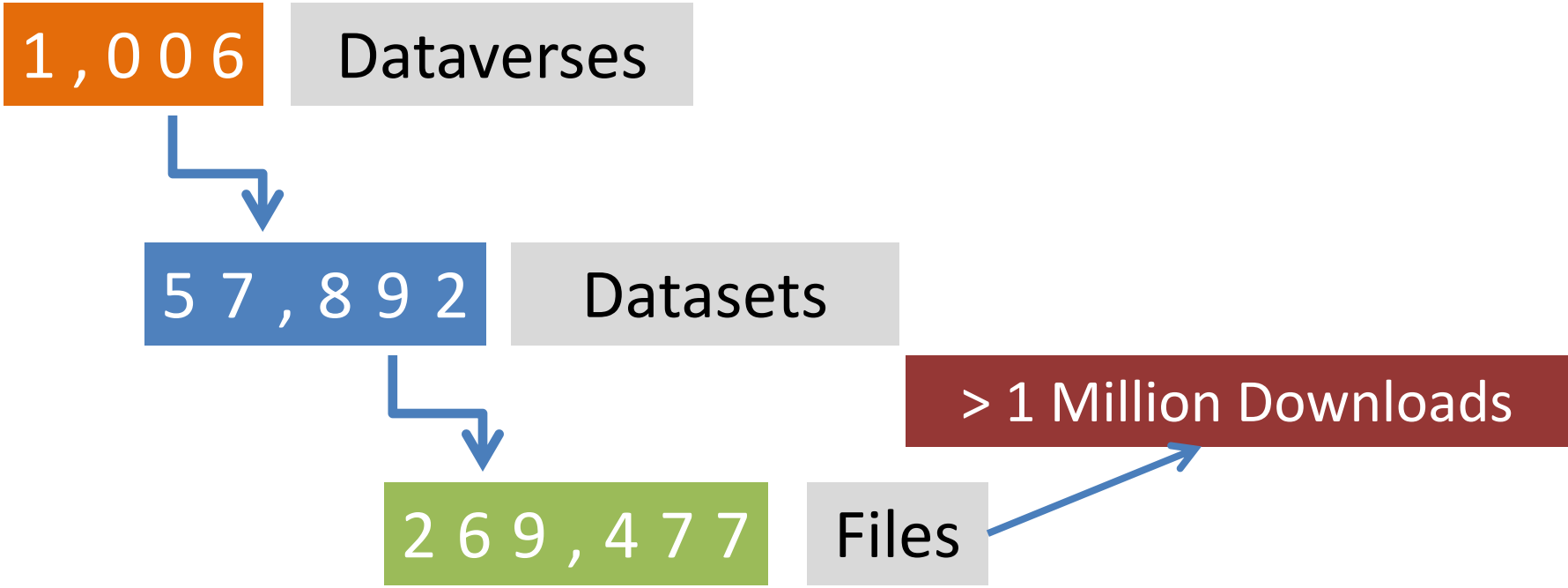
- Dataverse repository run at Harvard University



dataverse.harvard.edu

Harvard Dataverse

Open to all repository instance at Harvard currently has:



*number from April 25, 2015

The Dataverse Project

DATAVERSE BEST PRACTICES

Dataverse Best Practices (1)

- Standard Metadata Schemas
  - DDI & OAI DC
  - New in 4.0:
    - DataCite 3.1
    - ISA-Tab (biomedical)
    - VO Resource (astronomy)
    - DC Terms
  - Metadata can be exported in JSON & XML

Dataverse Best Practices (2)

- Metadata is always public once a dataset is published

- By default, datasets receive CC0 Waiver

- Even though default is CCO and we encourage open/public data, when needed, data files in a dataset can be made restricted, or terms of use can be added

Dataverse Best Practices (3)

- Formal Data Citation
  - Originally based off Altman + King 2007
  - Endorse + comply w/ 2014 Joint Declaration of Data Citation Principles (FORCE11)
    - Lead by Merce Crosas, Director of Data Science @ IQSS
  - Versioning and File Fixity
- Persistent IDs: DOI (DataCite/EZID)
  - Resolve to a dataset landing page, not directly to the data files

Data Citation Example

Principle 2: Credit and Attribution (e.g. authors, repositories or other distributors and contributors)

Principle 4: Unique Identifier (e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:** A persistent identifier that provides access and metadata

Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier

Principle 7: Specificity and verification
(e.g. the specific version used).
Versioning or timeslice information should be supplied with any updated or dynamic dataset.

Dataverse Best Practices (4)

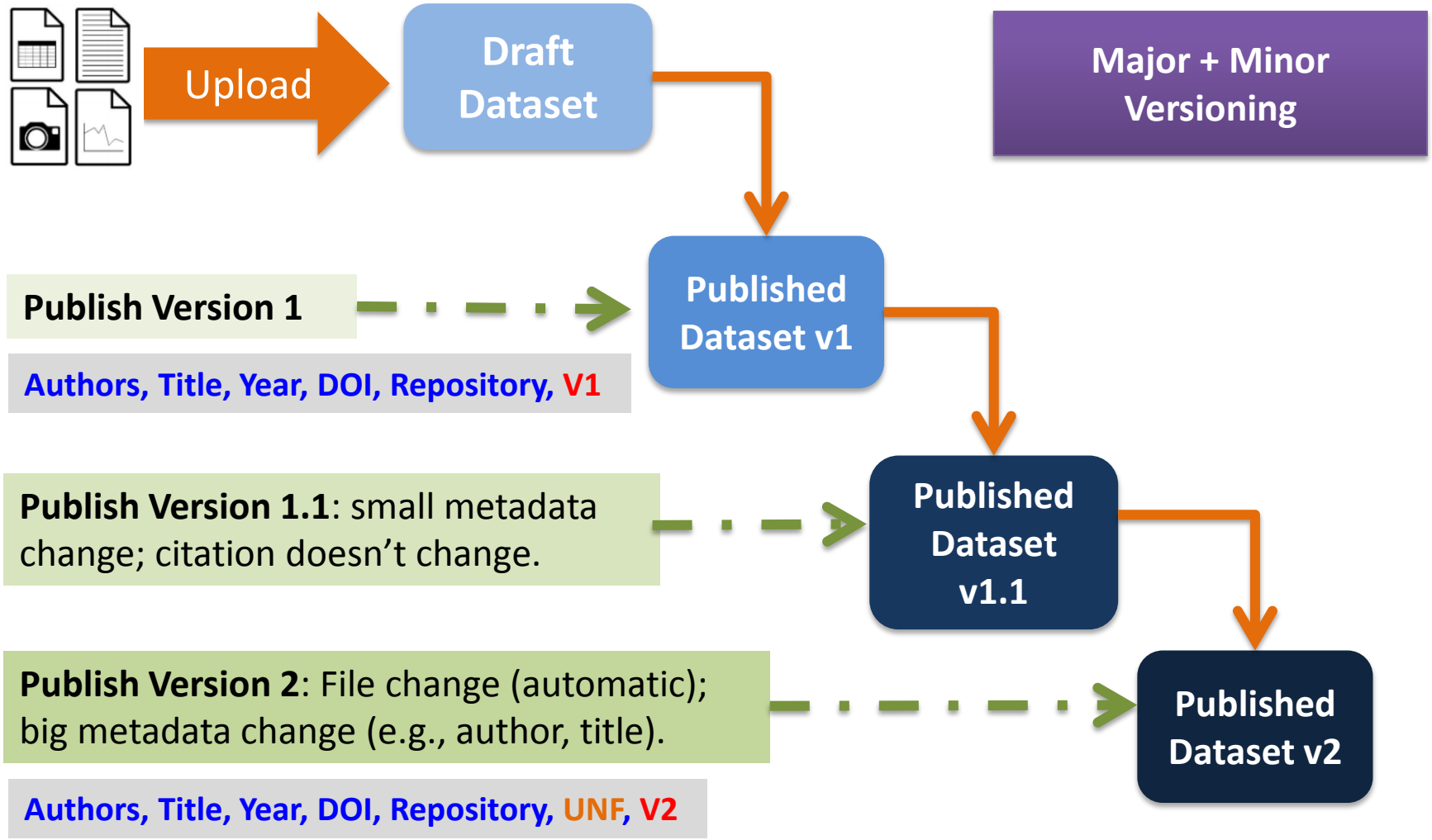
- Preservation format conversion for tabular data (extract column/variable metadata)
- File Fixity:
  - UNF (Altman, 2008) for tabular data
  - MD5 checksums for other files

Dataverse Best Practices (5)

- Data-PASS: (ICPSR, ODUM, NARA, ROPER,…)
  - Member of Data-PASS
- OAI-PMH: Harvesting metadata (DC, DDI)
  - From other Dataverse installations
  - From other OAI-DC compliant repositories
- If necessary: Deaccession a Dataset

PUBLISHING WITH DATAVERSE

Rigorous Data Publishing Workflows



Publishing a Dataset

Dataverse For testing only... About Support Contact Elizabeth Quigley 2

TCDL Demo Dataverse (Harvard University) Unpublished

Demo Dataverse > TCDL Demo Dataverse > **TCDL Demo Dataset**

Success! – The files for this dataset have been updated.

Metrics 0 Downloads Publish Edit

TCDL Demo Dataset Draft Unpublished

Quigley, Elizabeth, 2015, "TCDL Demo Dataset", http://dx.doi.org/10.5072/FK2/C7SKKC, Demo Dataverse, DRAFT VERSION [UNF:6:/KWlzLRBtSgA5Y2TgzB8xw==]
If you use these data, please add this citation to your scholarly resources. Why Cite?

Download Citation

Description	Dataset created for demo purposes for Texas Conference on Digital Libraries.
Subject	Arts and Humanities

Files Metadata Terms Versions

Download Upload + Edit Files

AK_2002.tab
Tabular Data - 35.4 KB - Apr 21, 2015 - 0 Downloads
Original File MD5: 4f82c6d25c7fbce90e7c16abfa92de1b;
18 Variables, 517 Observations - UNF:6:/KWlzLRBtSgA5Y2TgzB8xw==
Explore Download

HowToDeposit4-01.png
PNG Image - 90.5 KB - Apr 21, 2015 - 0 Downloads
MD5: 96104c15decdb3429cc586889e3cafb1;
Download

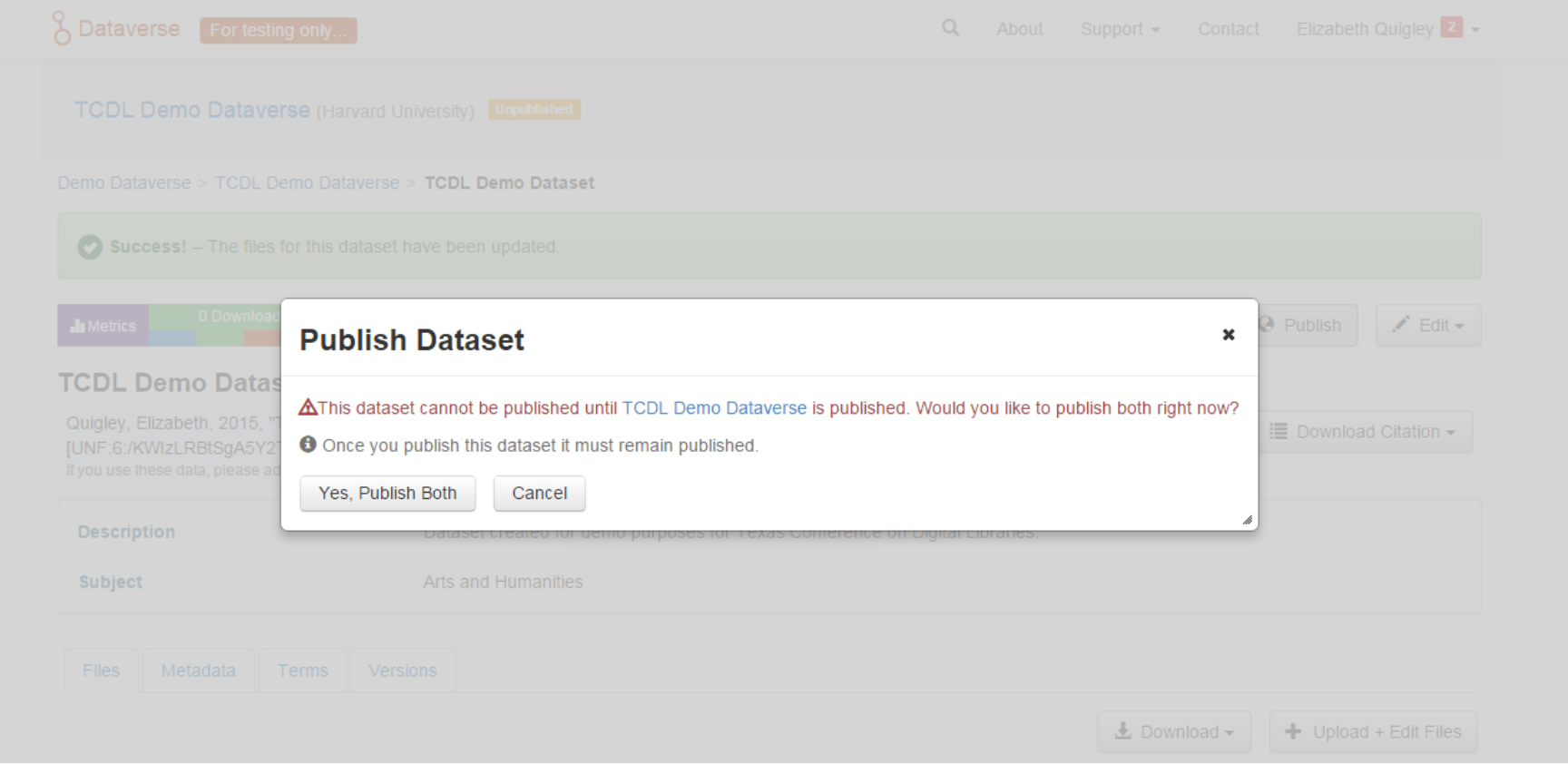
IQSSLogo-2014-Large (2).png
PNG Image - 108.5 KB - Apr 21, 2015 - 0 Downloads
MD5: 3db97461ea101a8f200fa6fccdd01808;
Download

l1.avge.fits
FITS - 7.7 MB - Apr 21, 2015 - 0 Downloads
MD5: 811f4d104ce96d34e66b552e8c84af48;
This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file:
INSTRUME; NAXIS0; NAXIS1; TELESCOP; DATE-OBS; CRVAL2; NAXIS; OBJECT; CD1_1; CRVAL1; EXPTIME;
Download

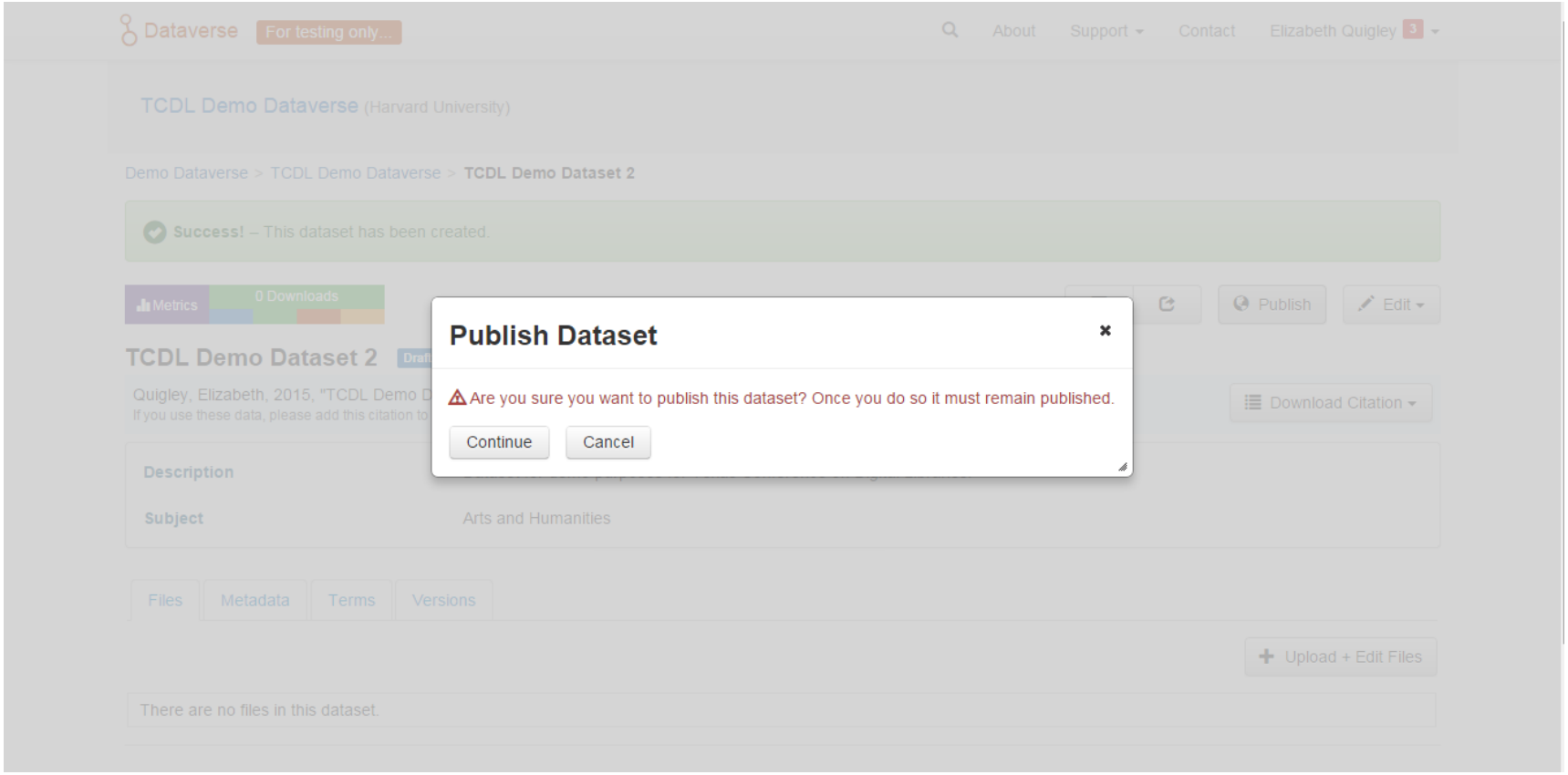
Data Science at The Institute for Quantitative and Social Science | Dataverse Project on | Code available at
Copyright © 2015, The President & Fellows of Harvard College | Privacy Policy

Powered by The Dataverse Project v. 4.0 build 1042

Publishing Dataverse + Dataset



Publishing a Dataset



The Dataverse Project

DISCOVERABILITY OF DATA

Searching for Data

- Search uses Solr, an open source search platform

- Solr is also used by:

NETFLIX

Smithsonian Institution

INTERNET ARCHIVE

Searching for Data

Dataverse

About | Support | Contact | Sign Up | Log In

Search this dataverse... | Find | Advanced Search | Add Data

- [x] Dataverses (1,008)
- [x] Datasets (57,805)
- [ ] Files (268,319)

Dataverse Category

Organization or Institution (29)
Researcher (25)
Journal (14)
Research Project (11)
Teaching Course (1)

Affiliation

Unknown (1,235)
Statistics Canada (1,172)
Statistique Canada (1,030)
United States Department of Commerce. Bureau of the Census (852)
United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics (676)

More...

Publication Date

2015 (4,207)
2009 (4,029)
2014 (2,695)
2007 (1,303)

1 to 10 of 58,813 results

Sort | « | < Previous | 1 | 2 | 3 | 4 | 5 | Next > | »

Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014
Apr 22, 2015 - MITx and HarvardX Dataverse
Hansen, John; Reich, Justin, 2015, "Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014", http://dx.doi.org/10.7910/DVN/29779, Harvard Dataverse, V3 [UNF:6:x50gdCP2OF7g789bpWm+kQ==]
[[NOTE: Placeholder for DOI. Data still in preparation. Data are currently only accessible to qualified reviewers.]] This dataset includes the home mailing addresses of all participants (registrants w...

Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh
Apr 22, 2015 - Randomized Controlled Trials in the Social Sciences Dataverse
Bryan, Gharad; Chowdhury, Shyamal; Mobarak, Ahmed Mushfiq, 2014, "Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh", http://dx.doi.org/10.7910/DVN/28277, Harvard Dataverse, V2
This paper presents the results of a financial intervention in northwestern Bangladesh intended to promote out-migration to nearby urban areas during the lean season before harvest in order to mitigat...

Replication Data for: Careful Commitments: Democratic States and Alliance Design
Apr 22, 2015 - The Journal of Politics Dataverse
Chiba, Daina; Johnson, Jesse C.; Leeds, Brett Ashley, 2015, "Replication Data for: Careful Commitments: Democratic States and Alliance Design", http://dx.doi.org/10.7910/DVN/DGHX1E, Harvard Dataverse, V1 [UNF:6:Scwy7ANl/KsVX9yelFi/aQ==]
Evidence suggests that leaders of democratic states experience high costs from violating past commitments. We argue that because democratic leaders foresee the costs of violation, they are careful to...

Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria
Apr 22, 2015
Gaskó, Noémi, 2015, "Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria", http://dx.doi.org/10.7910/DVN/USIP5E,

Searching for Data

Dataverse

About | Support | Contact | Sign Up | Log In

king | Find | Advanced Search | Add Data

- Dataverses (3)
- Datasets (451)
- Files (184)

Dataverse Category

Researcher (1)

Affiliation

Harvard University (60)
Unknown (10)
Puget Sound Transportation Panel (8)
UC San Diego (8)
CBS News (6)

More...

Publication Date

2007 (94)
2008 (70)
2015 (37)
2010 (19)
2014 (17)

More...

Author Name

King, Gary (108)
King County (Wash.) (103)
Adams, Greg (55)

1 to 10 of 638 results

Sort | « | < Previous | 1 | 2 | 3 | 4 | 5 | Next > | »

Gary **King** Dataverse (Harvard University)
Jan 12, 2007

King County, Washington : trails
Dec 12, 2011 - Harvard Geospatial Library Dataverse

King County (Wash.), 2011, "King County, Washington : trails"

... Location of trails in **King** County, Washington. City trails and private trails may not be complete ...
Producer Name: **King** County Parks GIS
Sampling Procedure: **King** County, Washington data was obtained via a CD-ROM. The data were in the format of ArcINFO
Author Name: **King** County (Wash.)

This dataset is harvested from our partners at Harvard Geospatial Library. Clicking the link will take you directly to the archival source of the data.

King County, Washington : streams
Dec 12, 2011 - Harvard Geospatial Library Dataverse

King County (Wash.); King County (Wash.). Water and Land Resources Division; Washington (State). Dept. of Ecology., 2011, "King County, Washington : streams"

... **King** County, Washington streams. Covers also portions of Snohomish and Pierce Counties. ...
Producer Name: **King** Co. Water and Land Resources
Sampling Procedure: **King** County, Washington data was obtained via a CD-ROM. The data were in the format of ArcINFO
Author Name: **King** County (Wash.)

This dataset is harvested from our partners at Harvard Geospatial Library. Clicking the link will take you directly to the archival source of the data.

King County, Washington : landmarks

Browsing for Data

Dataverse

About | Support | Contact | Sign Up | Log In

Search this dataverse | Find | Advanced Search | Add Data

- [x] Dataverses (1,008)
- [x] Datasets (57,805)
- [ ] Files (268,319)

Dataverse Category

Organization or Institution (29)
Researcher (25)
Journal (14)
Research Project (11)
Teaching Course (1)

Affiliation

Unknown (1,235)
Statistics Canada (1,172)
Statistique Canada (1,030)
United States Department of Commerce. Bureau of the Census (852)
United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics (676)

More...

Publication Date

2015 (4,207)
2009 (4,029)
2014 (2,695)
2007 (1,303)

1 to 10 of 58,813 results

Sort | « | < Previous | 1 | 2 | 3 | 4 | 5 | Next > | »

Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014
Apr 22, 2015 - MITx and HarvardX Dataverse
Hansen, John; Reich, Justin, 2015, "Socioeconomic Status Indicators of HarvardX and MITx Participants 2012-2014", http://dx.doi.org/10.7910/DVN/29779, Harvard Dataverse, V3 [UNF:6:x50gdCP2OF7g789bpWm+kQ==]
[[NOTE: Placeholder for DOI. Data still in preparation. Data are currently only accessible to qualified reviewers.]] This dataset includes the home mailing addresses of all participants (registrants w...

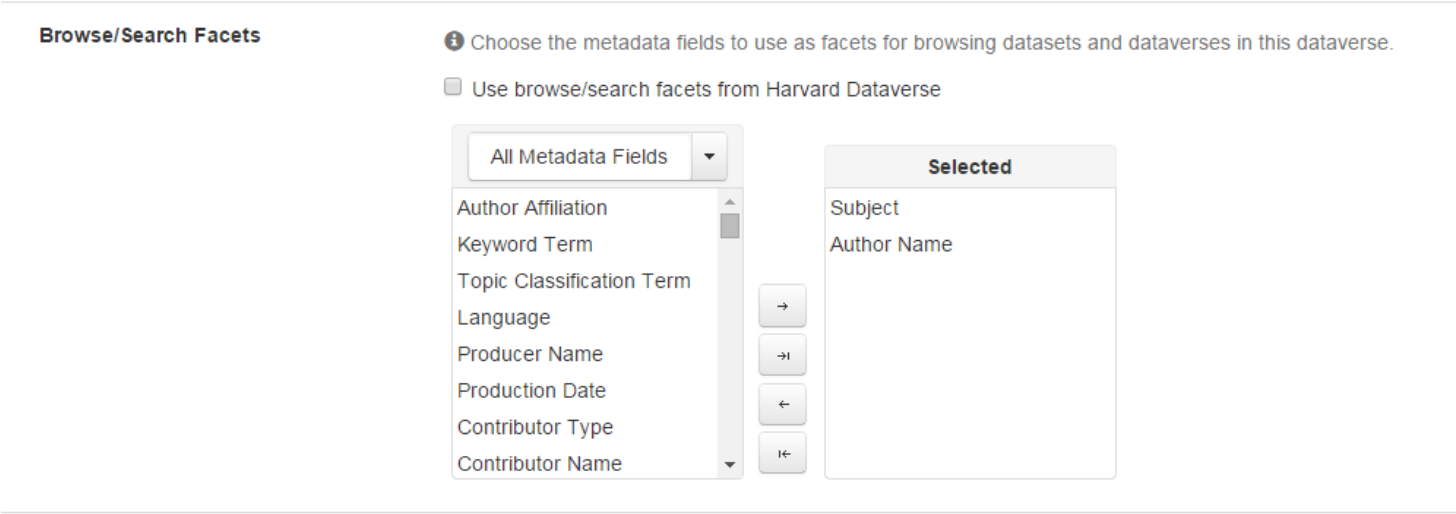
Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh
Apr 22, 2015 - Randomized Controlled Trials in the Social Sciences Dataverse
Bryan, Gharad; Chowdhury, Shyamal; Mobarak, Ahmed Mushfiq, 2014, "Underinvestment in a Profitable Technology: the Case of Seasonal Migration in Bangladesh", http://dx.doi.org/10.7910/DVN/28277, Harvard Dataverse, V2
This paper presents the results of a financial intervention in northwestern Bangladesh intended to promote out-migration to nearby urban areas during the lean season before harvest in order to mitigat...

Replication Data for: Careful Commitments: Democratic States and Alliance Design
Apr 22, 2015 - The Journal of Politics Dataverse
Chiba, Daina; Johnson, Jesse C.; Leeds, Brett Ashley, 2015, "Replication Data for: Careful Commitments: Democratic States and Alliance Design", http://dx.doi.org/10.7910/DVN/DGHX1E, Harvard Dataverse, V1 [UNF:6:Scwy7ANl/KsVX9yelFi/aQ==]
Evidence suggests that leaders of democratic states experience high costs from violating past commitments. We argue that because democratic leaders foresee the costs of violation, they are careful to...

Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria
Apr 22, 2015
Gaskó, Noémi, 2015, "Characterization and Detection of epsilon-Berge-Zhukovskii Equilibria", http://dx.doi.org/10.7910/DVN/USIP5E,

Browsing for Data

- All dataverses are able to select facets by going to the General Information option under the Edit button
- Facets available for all metadata domains supported in Dataverse



Thank you!

Any questions?

Contact: equigley@iq.harvard.edu

Learn more: dataverse.org

@dataverseorg

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).