New!

The Dataverse Project

@dataverseorg

Mercè Crosas, Director of Data Science, IQSS, Harvard University
@mercecrosas

Harvard Purdue Data Management Symposium, June 16-17, 2015

About Dataverse

Science requires community access to data

Technology Solution

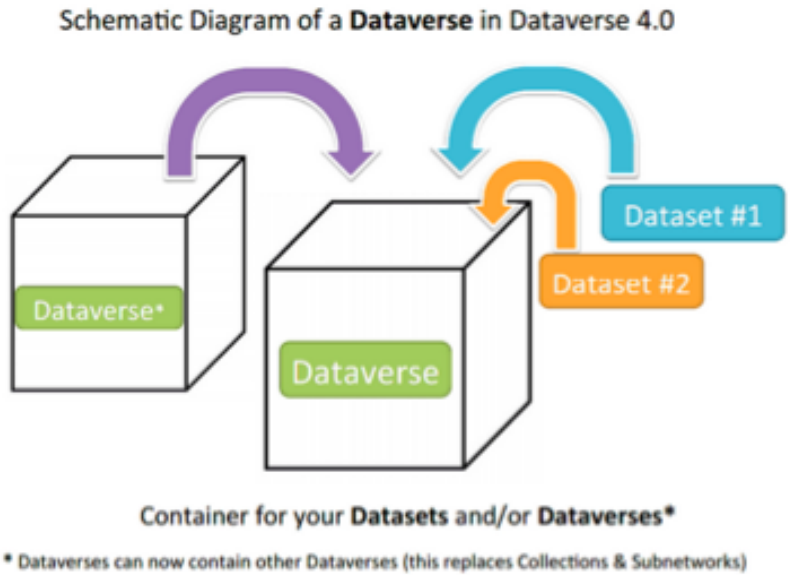
An open source software project for sharing, citing and archiving data

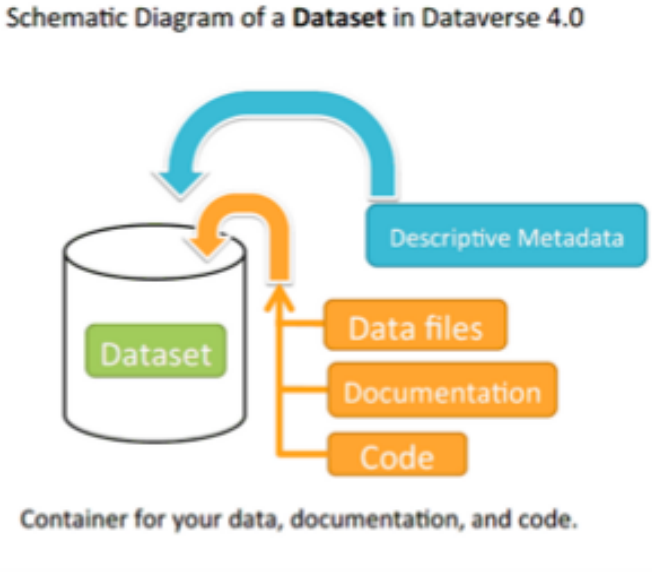
- Gives credit and control to data authors and distributors
- Follows best practices, standards for data management and archiving
- Dataverse development started in 2006 at Harvard's IQSS
- Now widely used, with a vibrant development and user community
- Helped instigate and is at the center of a cultural change toward open and reproducible research

In 2015 ...

Dataverse 4.0

A full rewrite that improves usability defines a rigorous and standardized data publishing workflow, and leverages the latest technologies.





Rich Set of Features

- Standard, persistent data citation
- Branding for each dataverse
- Standard, extensible metadata:
  - citation metadata
  - domain-specific metadata
  - file-level metadata
- Faceted search for all metadata
- Multiple levels of access control
  - CC0/ terms of use/ restricted
- Multiple roles and permissions
- Re-formatting of tabular data files
- Extraction of file metadata
- Versioning
- APIs for search, deposit, access

Upgraded Technology

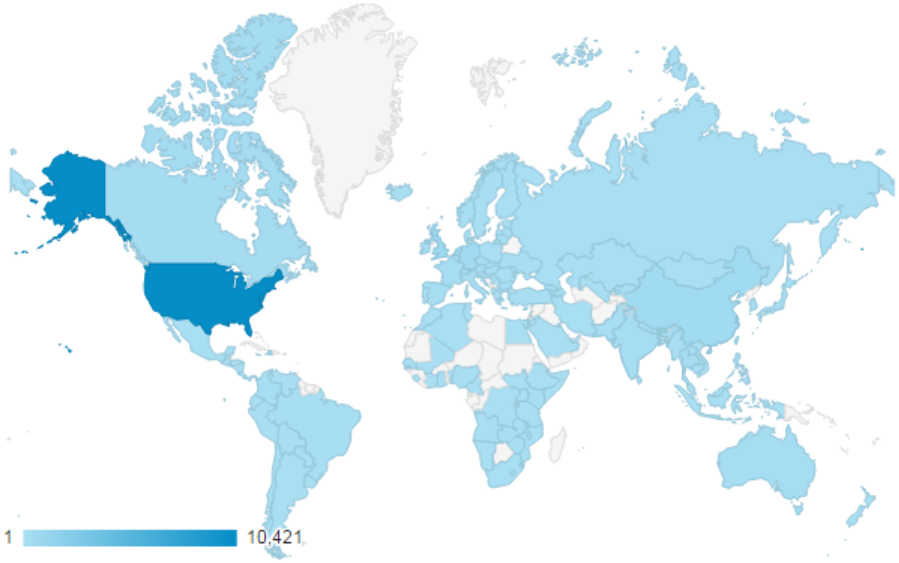
- UI improved by usability testing
- Built with open source solutions
- Enhanced UI framework
  - PrimeFaces and Bootstrap
- Widely used, community driven enterprise software platform
  - Java EE7 and Glassfish
- Reliable, scalable search platform
  - Solr
- Web standard programmatic interfaces
  - RESTful APIs
- Standards for archiving and interoperability
  - OAI-PMH, LOCKSS

Dataverse Installations worldwide

Dataverse software installations around the world serve as public data repositories (Harvard and ODUM Dataverses) or institutional research data repositories.

Harvard Dataverse

- A collaboration between the **Harvard Library** and IQSS
- Open to research data worldwide:
  - > 1000 dataverses
  - > 58,000 datasets
  - > 270,000 files
  - > 1.3 million downloads
  - > 11,000 registered users
- Includes dataverses for:
  - individual researchers
  - research teams
  - journals
  - institutions or organizations



- Rate of data deposit has increased by a factor of 30 since last year

Harvard Dataverse

A collaboration with Harvard Library, Harvard University IT, and IQSS

Metrics 1,326,251 Downloads

World Agroforestry Centre - ICRAF Dataverse

Population Services International (PSI) Dataverse

International Food Policy Research Institute (IFPRI) Dataverse

Murray Research Archive Dataverse

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

Dataverses (1,105)
Datasets (58,181)
Files (274,482)

Dataverse Category
Organization or Institution (94)
Research Project (76)
Researcher (52)
Journal (29)
Teaching Course (6)

Publication Date
2015 (13,609)
2007 (9,586)
2009 (6,251)
2014 (4,449)
2010 (4,195)

More...

Subject
Social Sciences (3,350)

1 to 10 of 59,286 Results

Sort « < Previous 1 2 3 4 5 Next > »

ICEWS Coded Event Data
Jun 16, 2015 - Integrated Crisis Early Warning System (ICEWS) Dataverse
Boschee, Elizabeth; Lautenschlager, Jennifer; O'Brien, Sean; Shellman, Steve; Starz, James; Ward, Michael, 2015, "ICEWS Coded Event Data", http://dx.doi.org/10.7910/DVN/28075, Harvard Dataverse, V2
Event data consists of coded interactions between socio-political actors (i.e., cooperative or hostile actions between individuals, groups, sectors and nation states). Events are automatically identif...

Replication data for: "Is More Better or Worse? New Empirics on Nuclear Proliferation and Interstate Conflict by Random Forests"
Jun 16, 2015 - Research & Politics Dataverse
Akisato Suzuki, 2015, "Replication data for: "Is More Better or Worse? New Empirics on Nuclear Proliferation and Interstate Conflict by Random Forests"", http://dx.doi.org/10.7910/DVN/29361, Harvard Dataverse, V2 [UNF:6:S2/p/BAsSUORBKxHmzbIgA==]
This contains the replication materials for the published version of the paper "Is More Better or Worse? New Empirics on Nuclear Proliferation and Interstate Conflict by Random Forests" in Research an...

Pheidole
Jun 16, 2015 - Pheidole Dataverse
Farnum, Charles, 2015, "Pheidole", http://dx.doi.org/10.7910/DVN/HMAFSD, Harvard Dataverse, V1
Partial list of Pheidole species for testing purposes.

ODAP

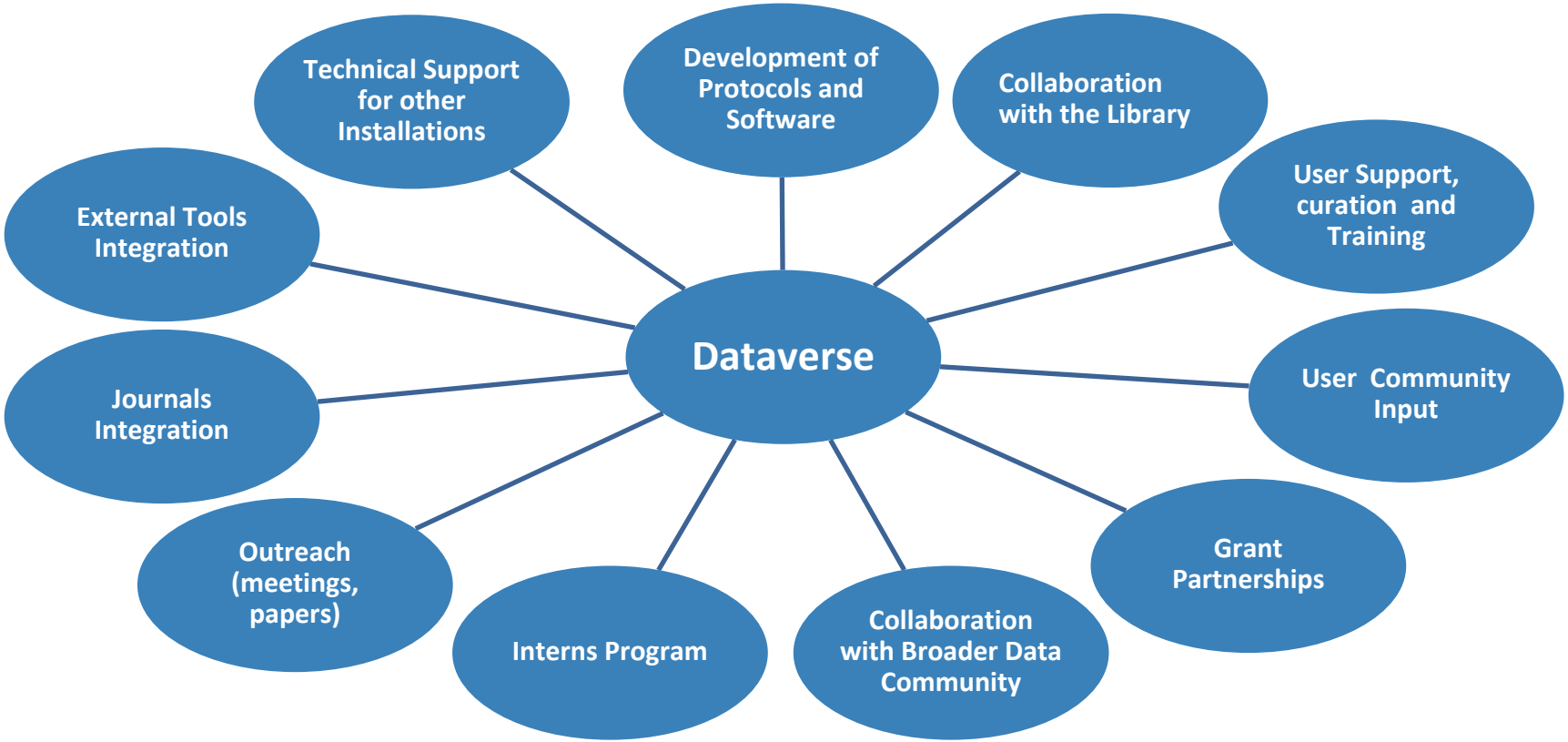
The Open Data Assistance Program at Harvard

Many Harvard researchers are subject to open-data policies from the journals publishing their articles or the agencies funding their research. Many others simply want to open up their data to realize the benefits of transparency, collaboration, data citation, research acceleration, and reproducibility. ODAP is a program to help them.

ODAP will offer advice and instruction on how to deposit data files in the Harvard Dataverse. When privacy is an issue, ODAP will offer advice on how to make data files as open as privacy constraints will allow. Since anyone in the world may deposit in Dataverse, ODAP's online assistance should help researchers everywhere. However, when online assistance isn't enough, ODAP staffers and volunteers will offer personal assistance to Harvard faculty, students, fellows, and postdocs. We encourage other institutions to offer personal assistance to their own researchers as well, and can work with them on how to do that.

If you're interested in providing open access to your data, you should also be interested in providing open access to the research articles reporting your analysis and conclusions. If you're at Harvard, we welcome your research publications, especially your scholarly articles, in our open-access repository, Digital Access to Scholarship at Harvard (DASH). For more details, see the Office for Scholarly Communication.

Dataverse is now more than a software project and a data repository



A vibrant community

Contributions:

- Internalization, translation to chinese (Fudan University)
- Integration with Archivematica (University of Toronto)
- Integration with iRODS (UNC)
- Integration with Shibboleth (Netherlands Dataverse, DANS)

What's next

- Support new data types
  - Sensitive information
  - Large-scale data (terabytes to petabytes)
  - Streaming data (e.g., sensors, cell phones)
- Make datasets shared in Dataverse more reusable
  - Extend APIs to build community of app contributors
  - Build integrated tools (e.g., data visualization, analysis)

dataverse demo

https://dataverse-demo.iq.harvard.edu

Thanks

@mercecrosas