The Dataverse Project

A Practical Example of Data Archiving

Eleni Castro > IQSS Harvard

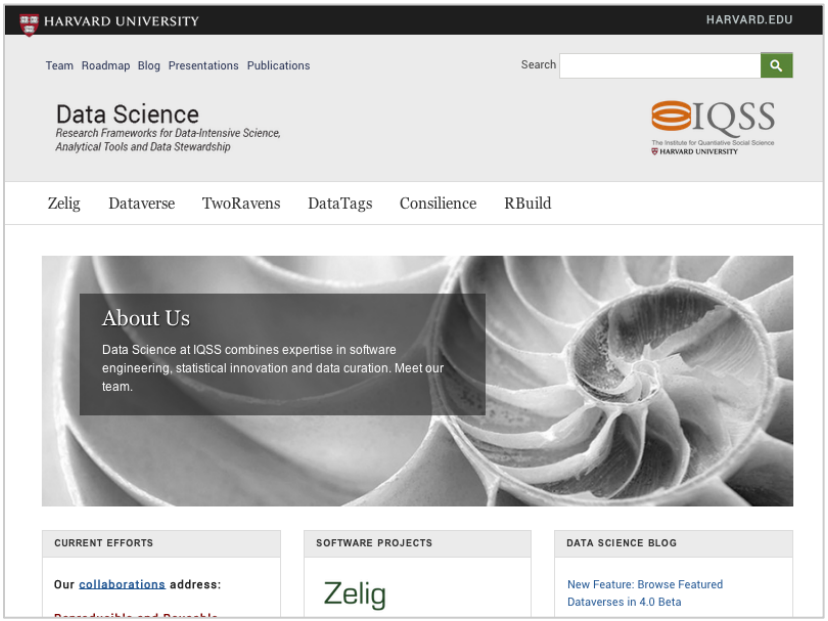
SHARE 2014 Fall Meeting

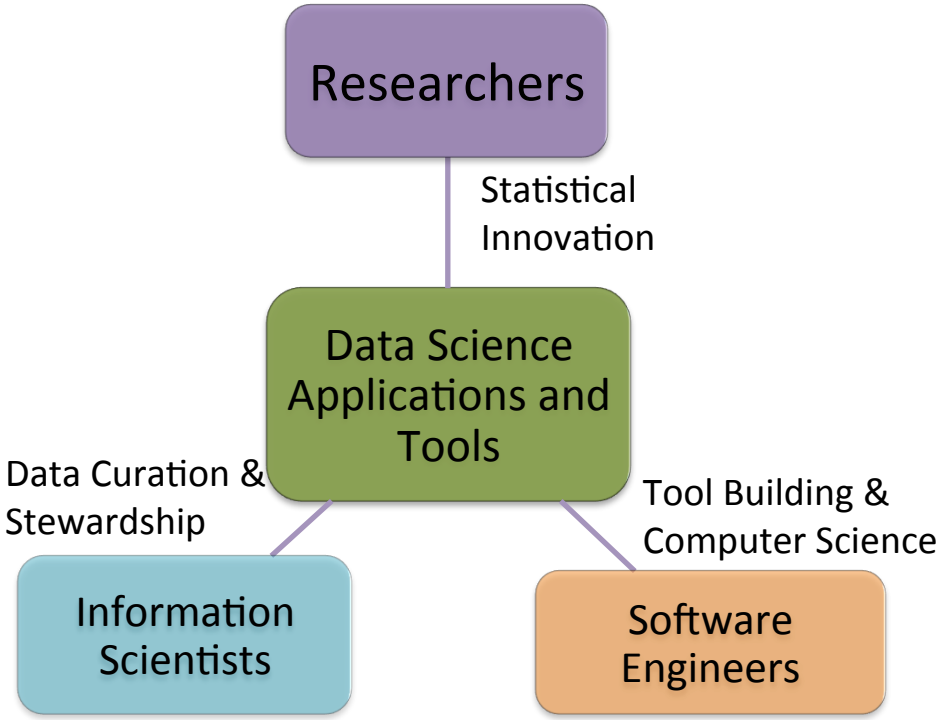
October 14, 2014

Image: https://flic.kr/p/7vu434s

Data Science Team

The Institute for Quantiative Social Science





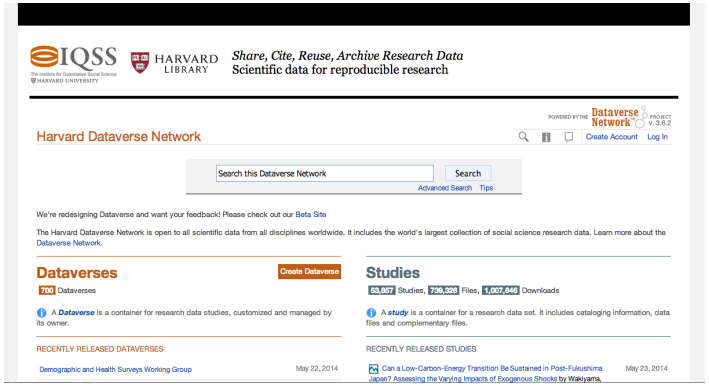
Find out more: http://datascience.iq.harvard.edu

Introduction to Dataverse

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

Provides incentives for researchers to share:

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



Harvard Dataverse (**open to all; general** repository instance at Harvard):

793 Dataverses

55,000 Datasets

750,703 Files

> 1 Million Downloads

The Dataverse Project

Why did we launch Dataverse?

1. Replication Standard (King, 1995)
2. Virtual Data Center (1999-2006)
3. The Dataverse Network® Project — Since 2006

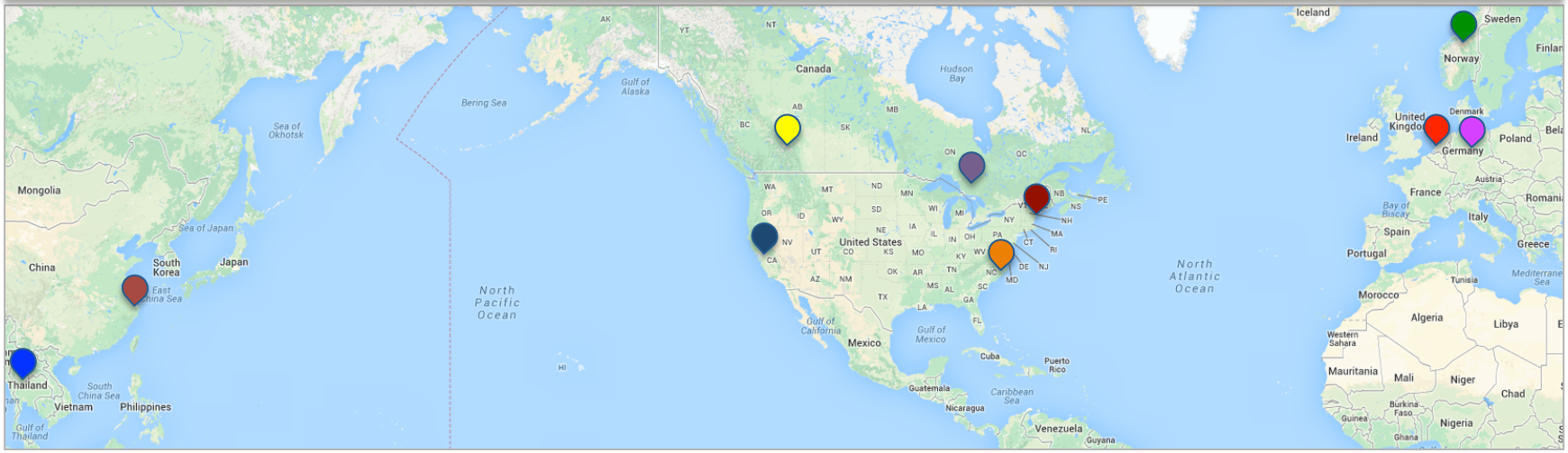
HARVARD LIBRARY

HMDC Harvard - MIT Data Center, Harvard University

IQSS The Institute for Quantiative Social Science

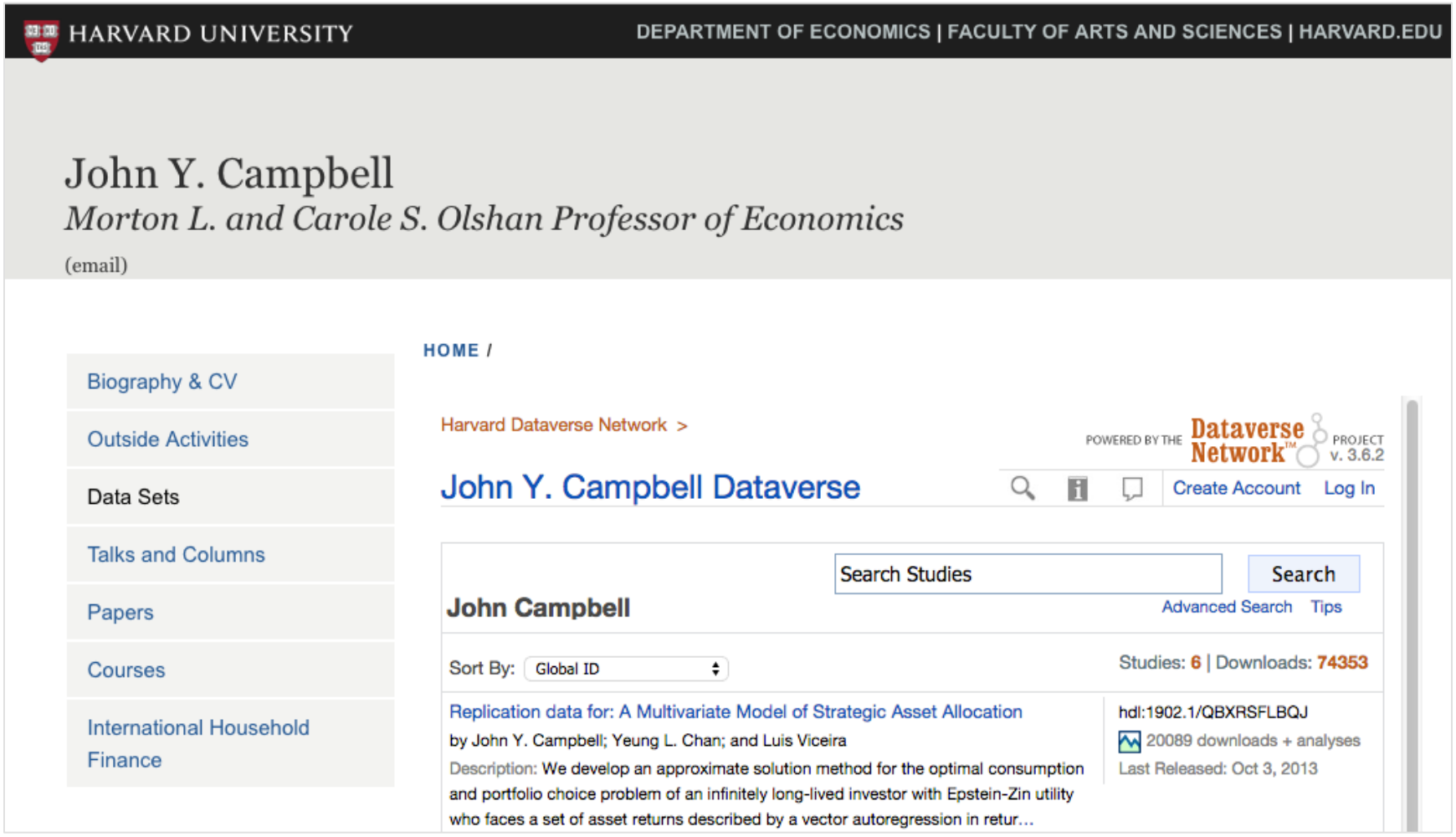
Who Uses Dataverse ?

Worldwide Dataverse Installations

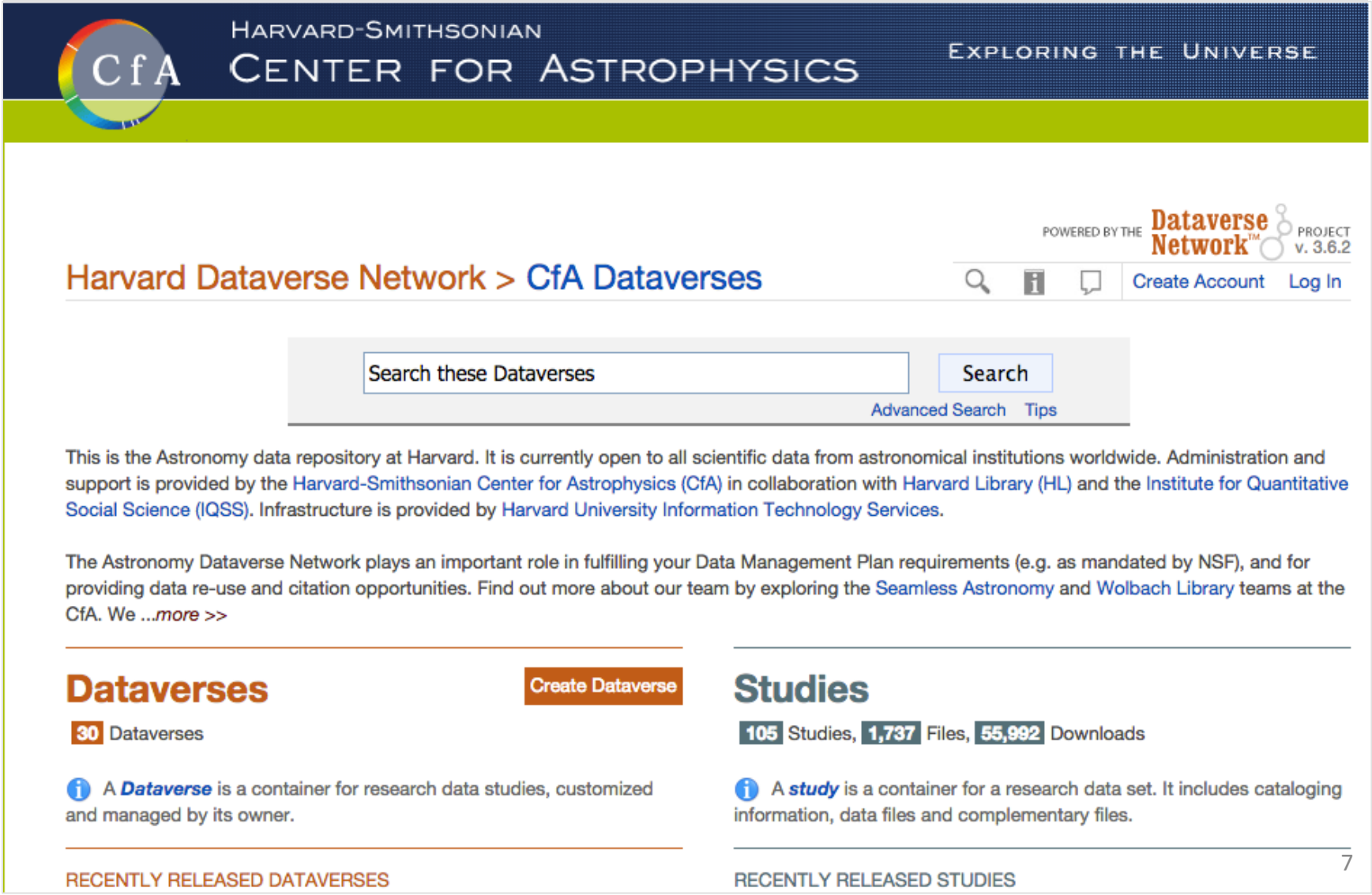


Institutions can setup/host their own Dataverse installation (UNC ODUM, Fudan Univ, Scholars Portal, DANS, etc) and within them can have dataverses for a variety of users (across all research domains): Researchers, Projects, Journals, etc.

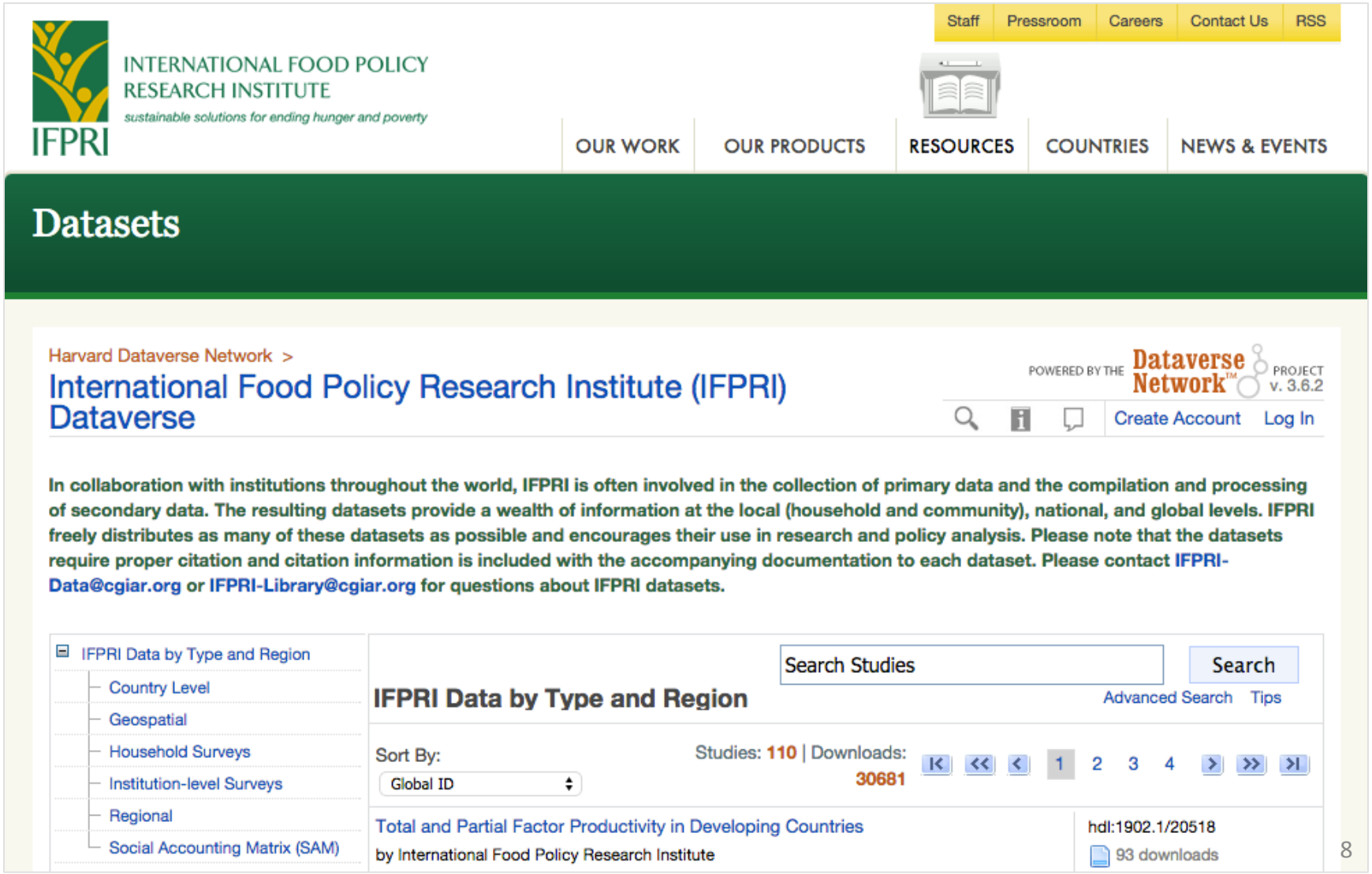
Example of a Scholar's Dataverse



Research Center Dataverse

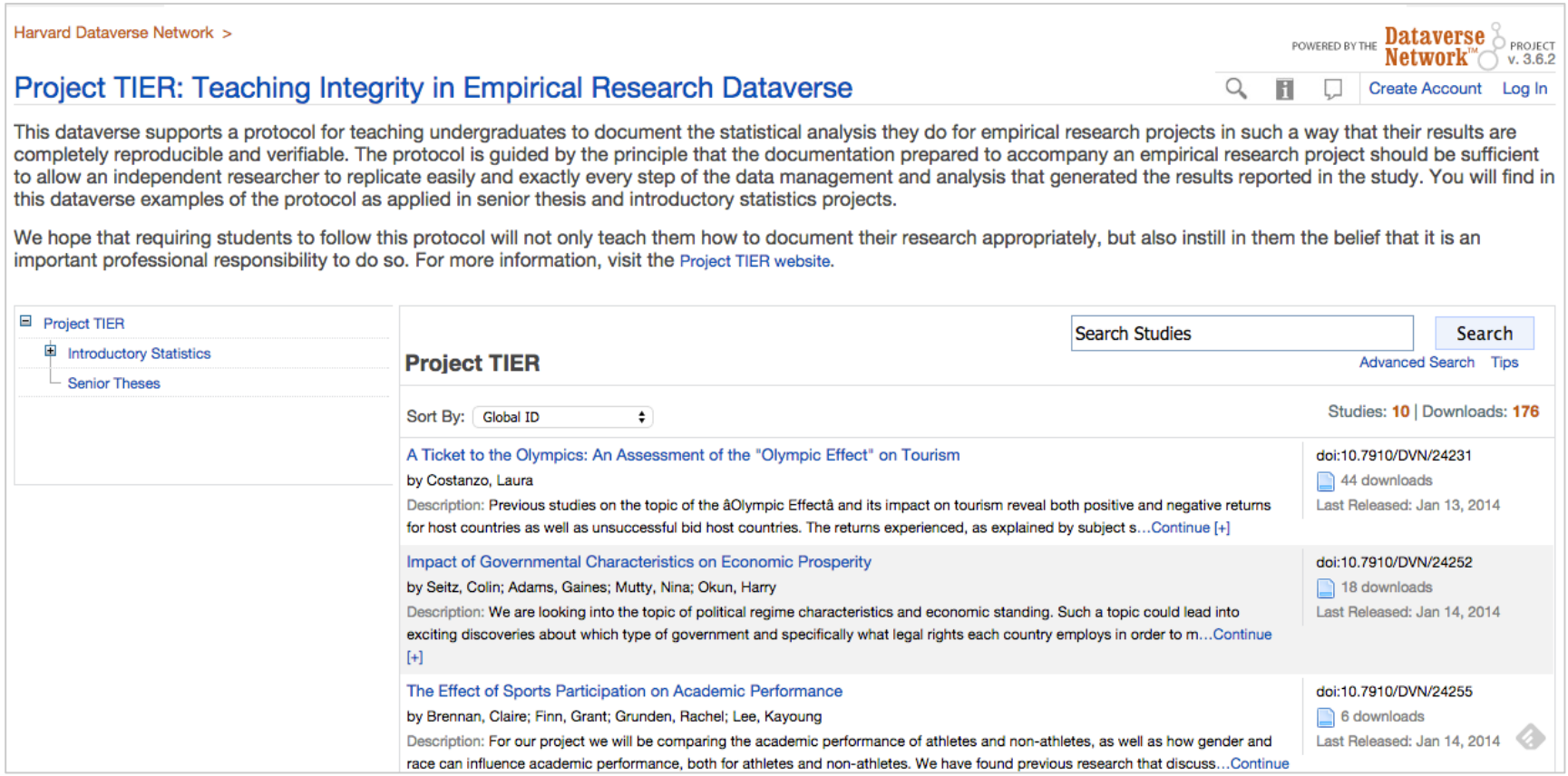


Example of an Institute Dataverse



Example of a Journal Dataverse

Dataverse for Teaching Replication



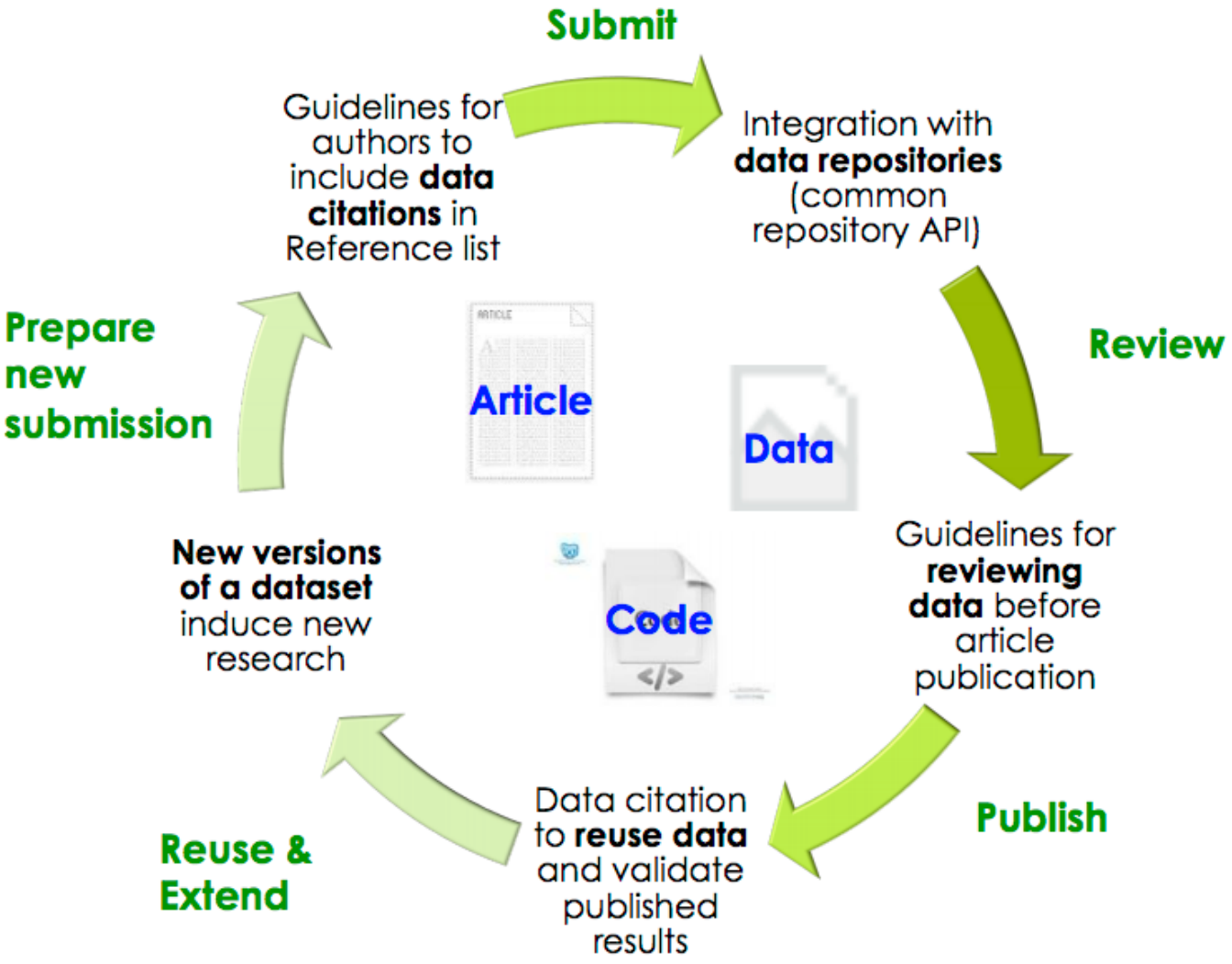
Dataverse Best Practices (1)

- Standard Metadata Schemas
  - DDI (great for social science data) & DC
  - Coming in 4.0: DataCite 3.0, ISA- Tab (biomedical), and VO Resource (astronomy)
- Formal Data Citation (Altman & King, 2007)
  - Endorse + comply w/ Joint Declaration of Data Citation Principles (incl. Crosas)
- Persistent IDs: Handles & DOI (DataCite/EZID)

Dataverse Best Practices (2)

- Fixity:
  - UNF (King & Altman) for tabular data
  - MD5 checksums for other files
- Open Data (+ metadata) Licenses (CC0) *or* custom
- **OAI-PMH**: harvesting metadata (DC, DDI,…)
- LOCKSS (replication of files) → Data-PASS

Towards An Integrated Publishing Workflow



API Integration with Dataverse

Data Deposit API (metadata + data w/ SWORDv2)

For depositing datasets into Dataverse via API

See: OJS-Dataverse Journal Integration Project
http://projects.iq.harvard.edu/ojs-dvn/home

Also: dvn R Package, **OSF** Dataverse Add-on, etc

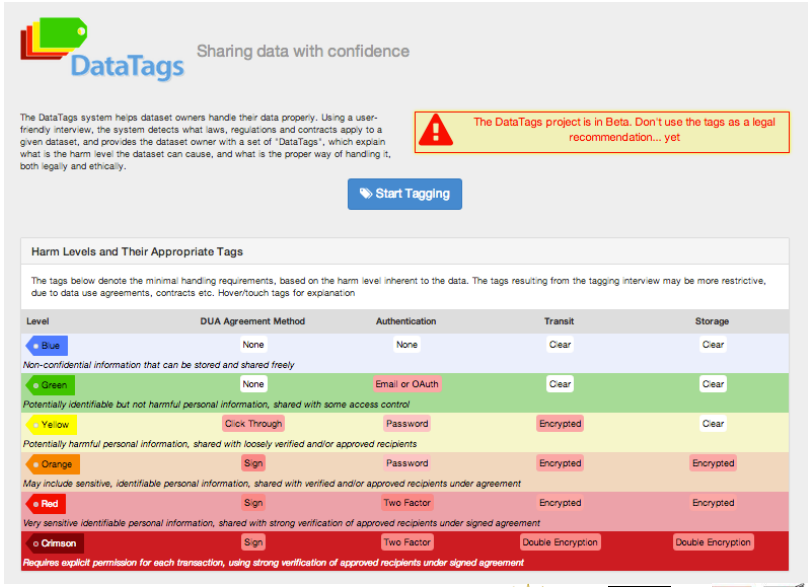
Data Sharing API

For searching/downloading Dataverse datasets (metadata + data) via API.

See: Thomas Leeper's dvn R package

Future of Dataverse?

- Dataverse 4.0 (*try* ***beta***)
  - Based on usability testing
- WorldMap Integration (geospatial viz w/ GeoConnect)
- **Sharing Privacy Sensitive Data**
  - Secure Dataverse
  - **DataTags** (questionnaires based on privacy laws)



Longer-Term

- Provenance Registry (data citation & provenance w/ SEAS (NSF))
- ORCID Integration (API)
- Large-scale datasets (efficient storage) → iRods w/ODUM
- Ensuring long-term preservation for more file formats (e.g., Archivematica)
- Integrate with more Publishing Systems

Thank you!

Contact: ecastro@fas.harvard.edu

More information: http://datascience.iq.harvard.edu/

Twitter: @thedataorg