The Dataverse Project

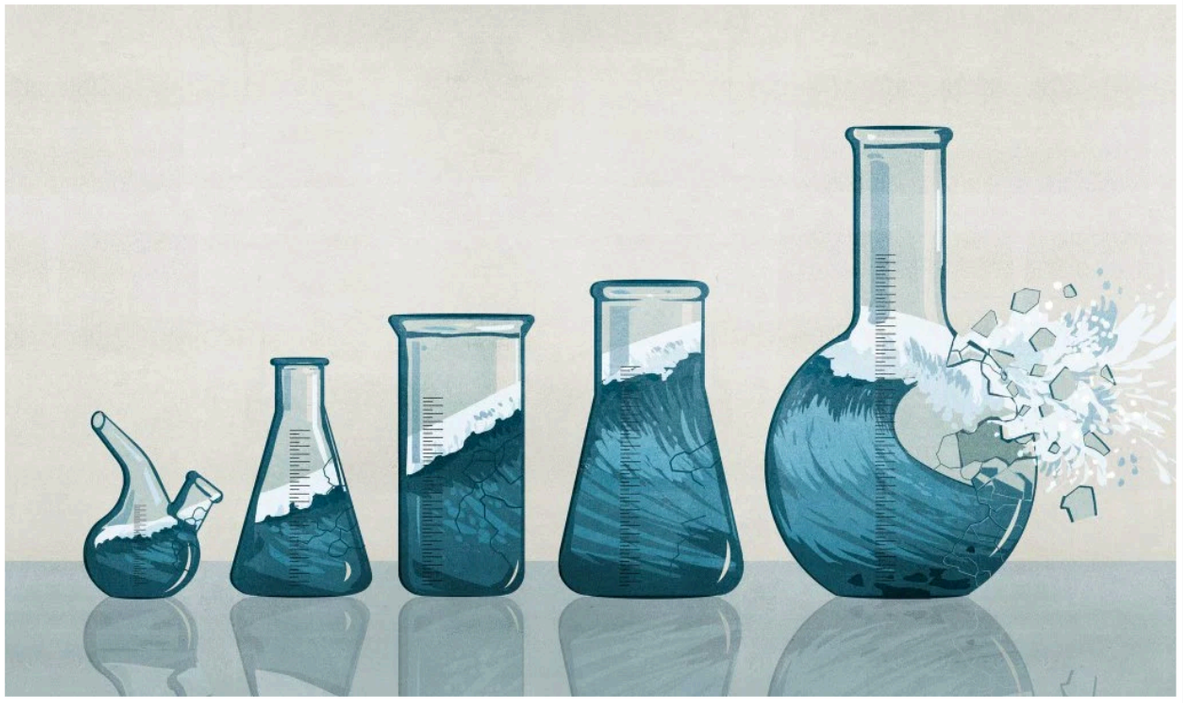
A data repository system for sharing and archiving research data

A Solution for Publishing FAIR research data: Findable, Accessible, Interoperable, Reusable

@dataverseorg @mercecrosas

Science isn't broken "It's just a hell a lot harder than we give it credit for"

FiveThirtyEight Science



THE SCIENTIFIC METHOD | 7:00 AM | AUG 19, 2015

Science Isn't Broken

It's just a hell of a lot harder than we give it credit for.

By CHRISTIE ASCHWANDEN
Graphics by RITCHIE KING

If you follow the headlines, your confidence in science may have taken a hit lately.Peer review? More like self-review. An investigation in November uncovered a scam in which researchers were rubber-stamping their own work, circumventing peer review at five high-profile

But it's self-correcting

INSIGHTS

Design principles for synthetic ecology p. 1425 ▸

Whacking hydrogen into metal p. 1429

SCIENTIFIC INTEGRITY

Self-correction in science at work

Improve incentives to support research integrity

By **Bruce Alberts,[1] Ralph J. Cicerone,[2] Stephen E. Fienberg,[3] Alexander Kamb,[4] Marcia McNutt,[5]* Robert M. Nerem,[6] Randy Schekman,[7] Richard Shiffrin,[8] Victoria Stodden,[9] Subra Suresh,[10] Maria T. Zuber,[11] Barbara Kline Pope,[12] Kathleen Hall Jamieson[13,14]**

POLICY

Week after week, news outlets carry word of new scientific discoveries, but the media sometimes give suspect science equal play with substantive discoveries. Careful qualifications about what is known are lost in categorical headlines. Rare instances of misconduct or instances of irreproducibility are translated into concerns that science is broken. The October 2013 *Economist* headline proclaimed "Trouble at the lab: Scientists like to think of science as self-correcting. To an alarming degree, it is not" (*1*). Yet, that article is also rich with instances of science both policing itself, which is how the problems came to *The Economist's* attention in the first place, and addressing discovered lapses and irreproducibility concerns. In light of such issues and efforts, the U.S. National Academy of Sciences (NAS) and the Annenberg Retreat at Sunnylands convened our group to examine ways to remove some of the current disincentives to high standards of integrity in science.

Like all human endeavors, science is imperfect. However, as Robert Merton noted more than half a century ago "the activities of scientists are subject to rigorous policing, to a degree perhaps unparalleled in any other field of activity" (*2*). As a result, as Popper argued, "science is one of the very few human activities—perhaps the only one—in which errors are systematically criticized and fairly often, in time, corrected" (*3*). Instances in which scientists detect and address flaws in work constitute evidence of success, not failure, because they demonstrate the underlying protective mechanisms of science at work.

Still, as in any human venture, science writ large does not always live up to its ideals. Although attempts to replicate the 1998 Wakefield study alleging an association between autism and the MMR (measles,

ILLUSTRATION: DAVIDE BONAZZI

Downloaded from www.sciencemag.org on July 24, 2015

Self-correction in science?

"trust, but verify"

Data (and code) should be shared for other researchers to inspect, test and reuse

U.S. Department of Health and Human Services National Institutes of Health

NIH *Data Science at NIH*

Data Science Community BD2K Commons News & Events

The Commons supports biomedical discovery by enabl
objec

Findable Accessible Interoperable and Reusable (FAIR)

Digital Object Compliance for the *Commons*

Data sharing is a key objective of the *Commons*, and cloud compu locate data and tools, as well as to make data (including the processed data that i reable by others. Access control can easily be implemented in the cloud so that d and tools can be appropriately and securely shared amongst groups authorized to use them, including the appropriate protecti s and access for human subjects data. Thus, while the cloud provides a computing environment to share data and tools in orde to be able to effectively use these digital objects, they must have attributes that make them **F**indable, **A**ccessible, **I**nteroperable and **R**eusable (FAIR). The *Commons* is intended to be a system that will do so.

A set of Digital Object Compliance principles that supports FAIR is currently under development. The Digital Object Compliance principles are expected to evolve over time as the ability to make digital objects meet the FAIR criteria increases, thereby improving the ability of digital objects to be shared and used more easily and effectively in the *Commons*.

To meet the most basic level of compliance, it is expected that digital objects would have the following elements:

- Unique digital object identifiers
- A minimal set of searchable metadata
- Physical availability through a cloud-based *Commons* provider
- Clear access rules and controls (especially important for human subjects data)
- An entry (with metadata) in one or more indices

SCIENTIFIC DATA

SCIENTIFIC DATA | COMMENT OPEN

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons ⊟ Show fewer authors

Research Data Lifecycle
(collect, process, analyze, compute)

Publish and Archive Data

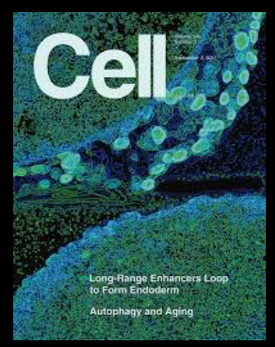
The Dataverse Project

Data Citation

Metadata

Access Controls

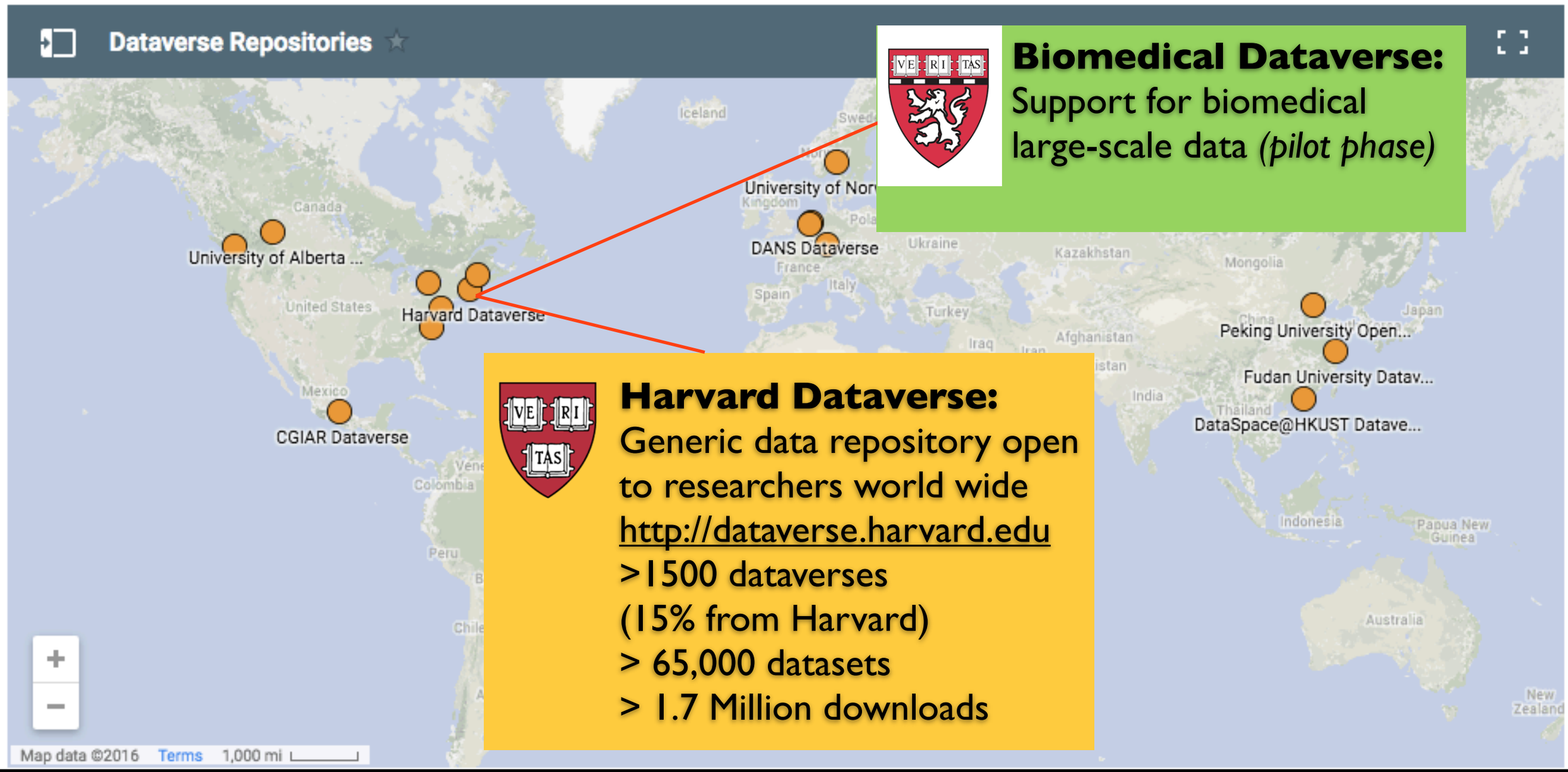
Publish Research Results



The Dataverse Project

http://dataverse.org

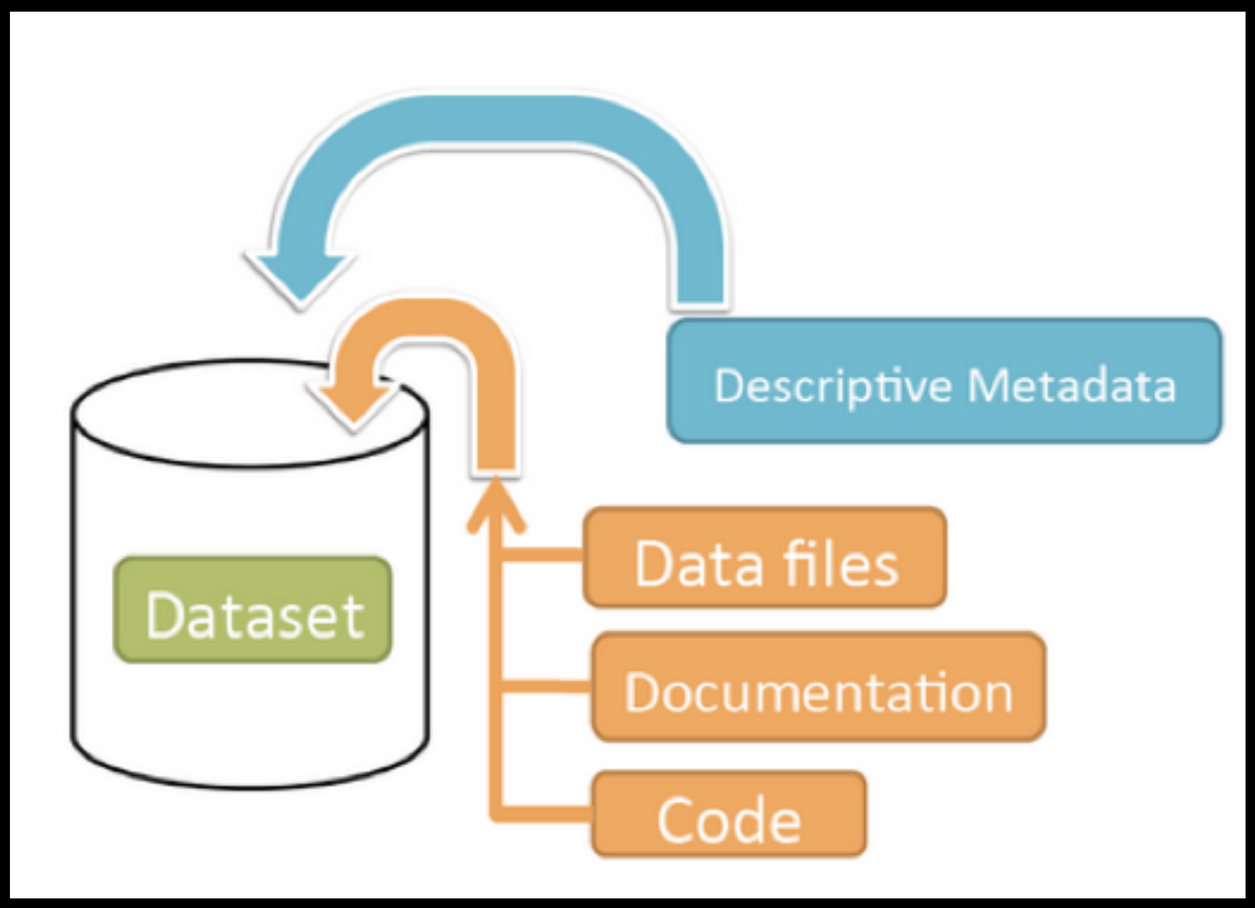
17 Installations | 1,500+ Dataverses | 65,000+ Datasets | 1,700,000+ Downloads



Biomedical Dataverse:
Support for biomedical large-scale data *(pilot phase)*

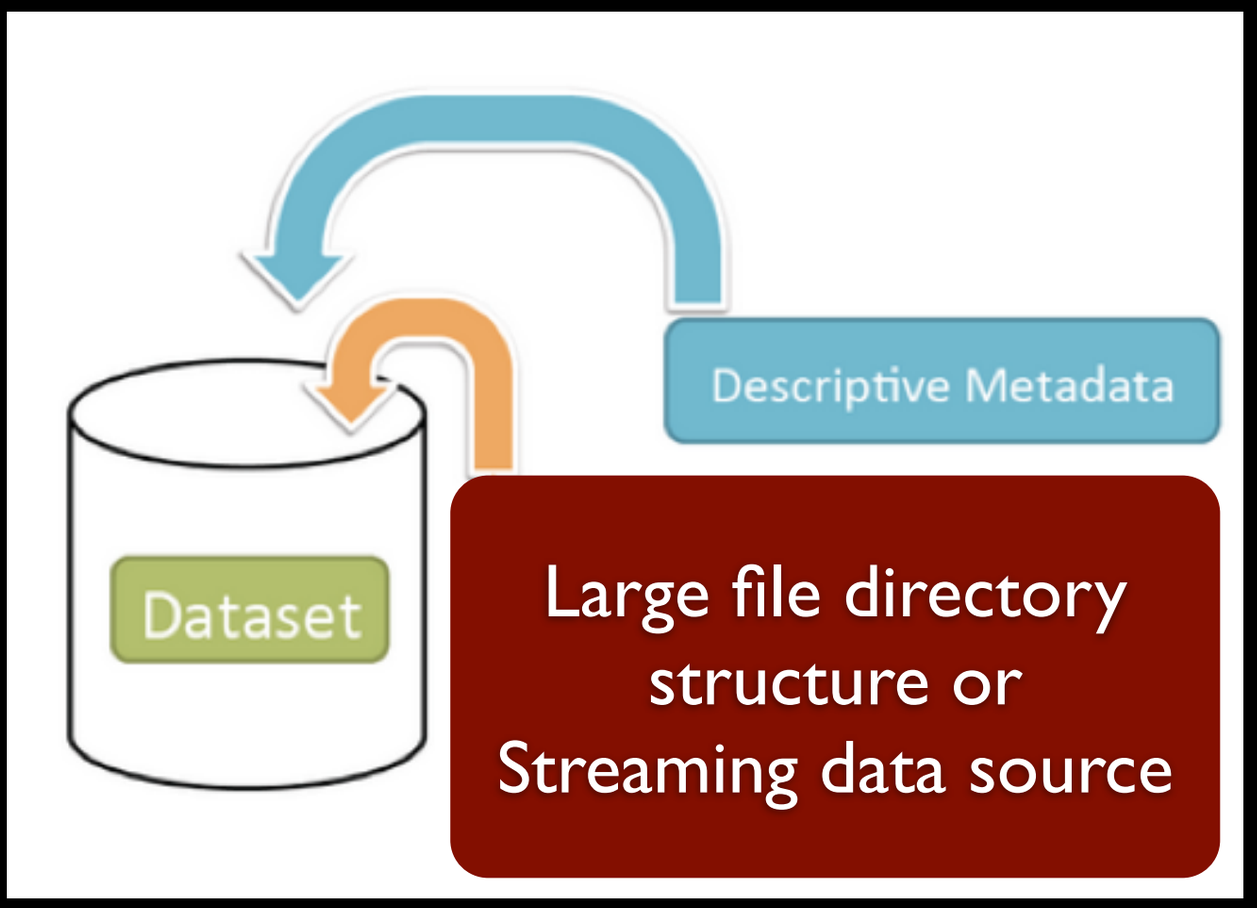
Harvard Dataverse:
Generic data repository open to researchers world wide
http://dataverse.harvard.edu
>1500 dataverses
(15% from Harvard)
> 65,000 datasets
> 1.7 Million downloads

Dataverse Now



Up to 100 files per dataset
Up to ~ Gb per file

Dataverse Big



> 1000s files per dataset
> Gb per file

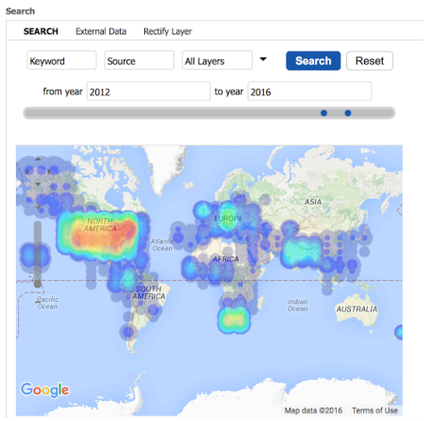
Dataverse Projects supporting Big Data

Biomedical Big Data

- 1000s image files per data set
- Non-http batch upload
- Data files in directory structure

SBGrid CONSORTIUM · IQSS · LINCS Library of Integrated Network-based Cellular Signatures · THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

Social Science Big Data



A Billion streaming Geotweets

IQSS · CENTER FOR GEOGRAPHIC ANALYSIS HARVARD UNIVERSITY · Alfred P. Sloan FOUNDATION

Cloud Dataverse

Massachusetts Open Cloud

The Dataverse Project

+ Swift Storage (scalable, access to computing)

Integration with Data Management Systems

STARFISH + The Dataverse Project at HMS

iRODS + The Dataverse Project at UNC