A data repository system for sharing and archiving research data

A Solution for Publishing FAIR research data: Findable, Accessible, Interoperable, Reusable
Science isn’t broken
“It’s just a hell a lot harder than we give it credit for”

But it’s self-correcting

**Science Isn’t Broken**

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

By Christie Aschwanden

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 smear in which researchers were rubber-stamping their own work, circumventing peer review at five high-profile nonprofits.

Week after week, news outlets carry word of new scientific discovery, but the media sometimes give 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 American Association for the Advancement of Science (AAAS) launched an initiative 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—which are not the only one—in which errors are systematically criticized and finally either, in time, corrected” (3). Instances in which scientists detect and address flaws in their work constitute evidence of success, not failure, because they demonstrate the underlying protective mechanisms of science at work.

Still, as in any human venture, errors will arise. Science will not always live up to its ideals. Although stamps to replace the 1998 Waterfield study alleging an association between autoxen and the MMR vaccine.
Self-correction in science?

“trust, but verify”

Data (and code) should be shared for other researchers to inspect, test and reuse
Findable Accessible Interoperable and Reusable (FAIR)

The Commons supports biomedical discovery by enabling the sharing of digital objects. Data sharing is a key objective of the Commons, and cloud computing provides an ideal environment to share and disseminate data and tools, as well as to make data (including the processed data that is part of research) and tools accessible to others. Access control can easily be implemented in the cloud so that data and tools can be appropriately and securely shared amongst groups that are authorized to use them, including the appropriate protections and access for human subjects data. Thus, while the cloud provides a computing environment to share data and tools in order to be able to effectively use these digital objects, they must have attributes that make them Findable, Accessible, Interoperable and Reusable (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
The FAIR Guiding Principles for scientific data management and stewardship

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

The Dataverse Project

Publish and Archive Data

Data Citation → Metadata → Access Controls

Publish Research Results

The Dataverse Project

Publish and Archive Data

Data Citation → Metadata → Access Controls

Publish Research Results

The Dataverse Project
Harvard Dataverse:
Generic data repository open to researchers worldwide
http://dataverse.harvard.edu
> 1,500 dataverses
(15% from Harvard)
> 65,000 datasets
> 1.7 Million downloads

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

The Dataverse Project
http://dataverse.org

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

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

Dataverse Big

- > 1000s files per dataset
- > Gb per file

Large file directory structure or Streaming data source
Dataverse Projects supporting Big Data

Biomedical Big Data

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

Social Science Big Data

A Billion streaming Geotweets

Cloud Dataverse

- Swift Storage
  (scalable, access to computing)

Integration with Data Management Systems

- STARFISH +
- iRODS +