Best Practices
Data should be Findable, Accessible, Interoperable, Reusable (FAIR) by machines

To be Findable:
- global, persistent ID
- registered, indexed

To be Accessible:
- open, standard protocol
- open metadata

To be Interoperable:
- references to other metadata
- FAIR vocabularies

To be Reusable:
- standard, rich metadata
- clear data licenses
- provenance

“Data Authorship as an Incentive to Data Sharing”
Today’s Bibliographies and CVs


Altman, M., Crosas, M. 2013. The Evolution of Data Citation: From Principles to Implementation. IASSIST Quarterly. p. 62

Future Bibliographies and CVs

data sets


Repositories should implement data citation maximizing discovery and access

Required:
• persistent ID/url resolves to dataset landing page

Recommended:
• landing page includes human- and machine-readable metadata

Optional:
• content negotiation for more accessible metadata

Fenner et al, 2016, “A Data Citation Roadmap for Scholarly Repositories” BioArxiv (preprint)
Synthesis Group, 2014, Joint Declaration of Data Citation Principles (Force11)
Dataverse
An open-source platform for building any type of data repository

With a growing community of developers and users

http://dataverse.org
Dataverse implements key aspects needed in modern data repositories

- Incentives to share data: citation, control, branding
- Citation to each version of the data with DOI or Handle
- Standard metadata for Discoverability
- Tiered access to non-public data
- Support for guestbook, data user agreements, licenses
- Commitment to data archival & preservation
Other features that make Dataverse a broad solution

- Workflows to publish data, including review and curation
- Roles and permissions to support data publishing and access
- API for search, deposit, and access of metadata and data
- Customization for each Dataverse
- Extension of metadata with custom metadata blocks
- Extraction of metadata from data files:
  - variables, observations, and summary stats from tabular data files
  - observation metadata from FITS astronomy files
  - GIS metadata from geospatial data files
Challenges
Current Challenges for Dataverse Repositories

- Datasets have to be small
- Hard to copy 40 PB over the internet
- Not every one has the right compute infrastructure

DATA REPOSITORIES NEED CLOUDS
CLOUDS NEED DATA REPOSITORIES
Cloud Dataverse on Massachusetts Open Cloud

Data depositor

Data users

Swift
Object Storage

Giji
Horizon

Sahara
Compute

Nova
Analytics

Compute

Compute
World Wide Data Federation

22 Installations
2,133 Dataverses
18,690 Datasets
2,400,322 Downloads
In order to build data sharing incentives, all parties need to be on board.

- **Publishers & Journals**: connect articles to data
- **Bibliographic repositories**: count data citations
- **Discovery Indexes & Registries**: find data
- **Researchers**: deposit data, get credit
- **Funders**: open data
- **Institutions**: support data
Recommendations
Rise of data-centric computation requires attractive ways to bring data to the cloud, and avoid moving the data back and forth.

Cloud computing should be integrated with data repositories (such as the Dataverse open-source repository platform).

To help establish best practices in the research workflow, we need to:

- Integrate seamlessly the entire pipeline: from data collection and analysis, to data sharing and archiving.
- Track documentation/metadata about the data through the cycle.
- Track data user agreements, licenses, and security restrictions that go with the data.
- Support tiered-access and tiered-use.