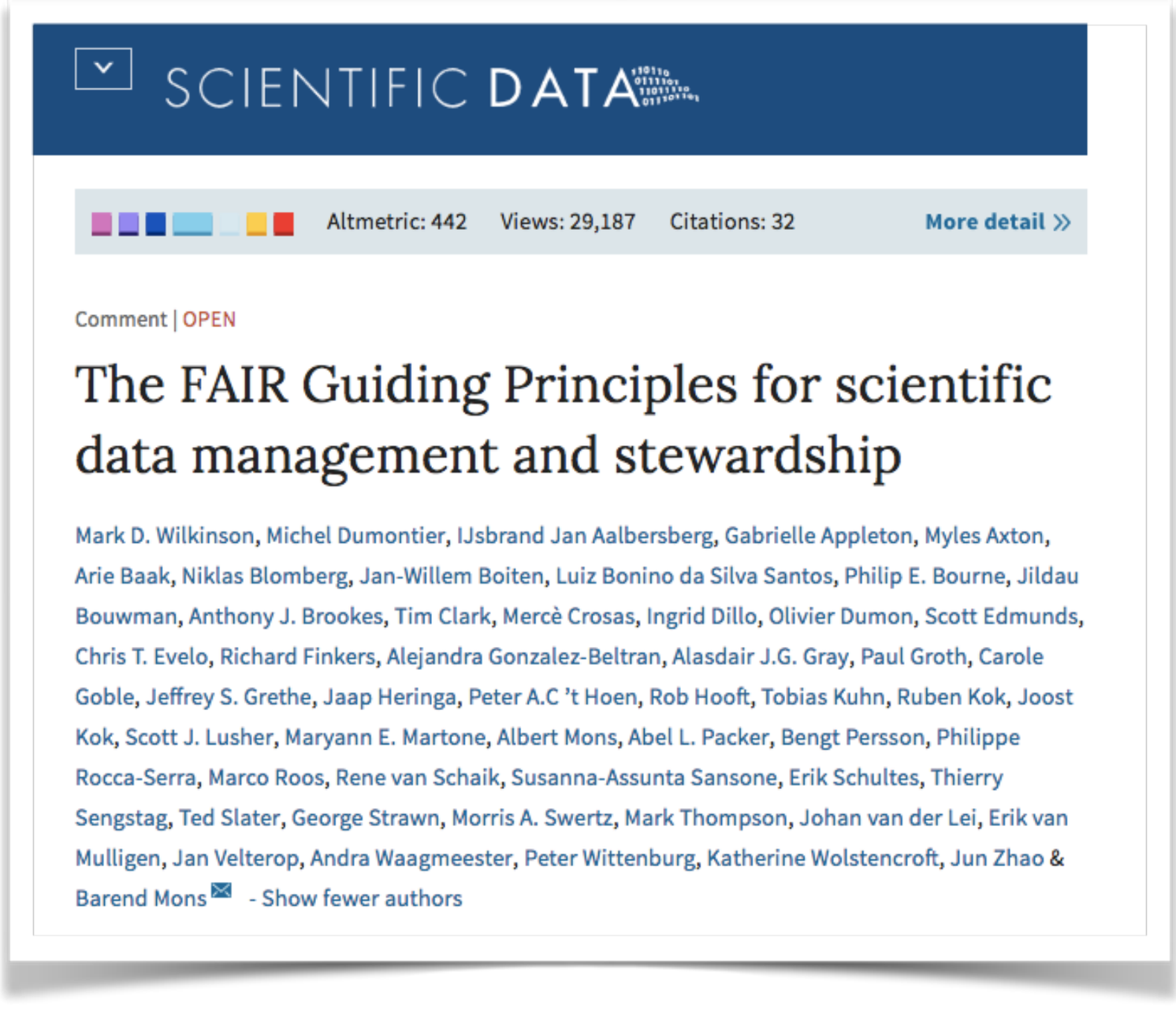
The Dataverse® Project

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science (IQSS)
Harvard University
@mercecrosas mercecrosas.com

Open Research Cloud, May 11, 2017

Best Practices

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



SCIENTIFIC DATA

Altmetric: 442 Views: 29,187 Citations: 32 More detail »

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

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

Wilkinson et al, 'The FAIR Guiding Principles scientific data management and stewardship," Nature Scientific Data, 2016;

“Data Authorship as an Incentive to Data Sharing”

The NEW ENGLAND JOURNAL of MEDICINE

SOUNDING BOARD

Data Authorship as an Incentive to Data Sharing

Barbara E. Bierer, M.D., Mercè Crosas, Ph.D., and Heather H. Pierce, J.D., M.P.H.

N Engl J Med 2017; 376:1684-1687 | April 27, 2017 | DOI: 10.1056/NEJMsb1616595

Today's Bibliographies and CVs

data sets

Future Bibliographies and CVs

Sweeney L, **Crosas M**, Bar-Sinai M. 2015, "Sharing Sensitive Data with Confidence: the DataTags System" Journal of Technology Science

Crosas, M, King, G, Honaker, J, Sweeney, L, 2015 "Automating Open Science for Big Data" The ANNALS of the American Academy of Political and Social Science, volume 659

Altman M, Castro E, **Crosas M**, Durbin P, Garnett A, Whitney J. 2015, "Open Journal Systems and Dataverse Integration-- Helping Journals to Upgrade Data Publication for Reusable Research" Code4Lib Journal, Issue 30

Starr J, Castro E, **Crosas M**, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. 2015, "Achieving human and machine accessibility of cited data in scholarly publications" PeerJ Computer Science, 1:e1 https://dx.doi.org/10.7717/peerj-cs.1

Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer,K., **Crosas, M.,** Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D.W., Kashyap,V., Mahabal, A., Siemiginowska, A., Slavkovic, A., 2014. 10 Simple Rules for the Care and Feeding of Scientific Data, PLoS Comput Biol, doi:10.1371/journal.pcbi.1003542

Pepe, A., Goodman, A., Muench, A., **Crosas, M.,** Erdmann, C., 2014. How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE, DOI: 10.1371/journal.pone.0104798

Crosas, M., 2013. A Data Sharing Story, Journal of eScience Librarianship, 2013, 1(3), 173-179, http://dx.doi.org/10.7191/jeslib.2012.1020

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

Ansolabehere, Stephen; Ban, Pamela; Snyder, James M., Jr., 2017, "State Legislative Historical Elections", doi:10.7910/DVN/LEMNXZ, Harvard Dataverse, V1, UNF:6:8UQYfDIsmII/tgD+Hrv/8Q==

King, Gary; Pan, Jennifer; Roberts, Molley, 2013, "Replication data for: How Censorship in China Allows Government Criticism but Silences Collective Expression", doi:10.7910/DVN1/22691, Harvard Dataverse, V4

Stephen Ansolabehere; Jonathan Rodden, 2011, "Colorado Data Files for State Legislative Elections", hdl:1902.1/15385, Harvard Dataverse, V2, UNF:5:jNUA7tB3bFeMcC2oGBvdHw==

Sweeney L, **Crosas M**, Bar-Sinai M. 2015, "Sharing Sensitive Data with Confidence: the DataTags System" Journal of Technology Science

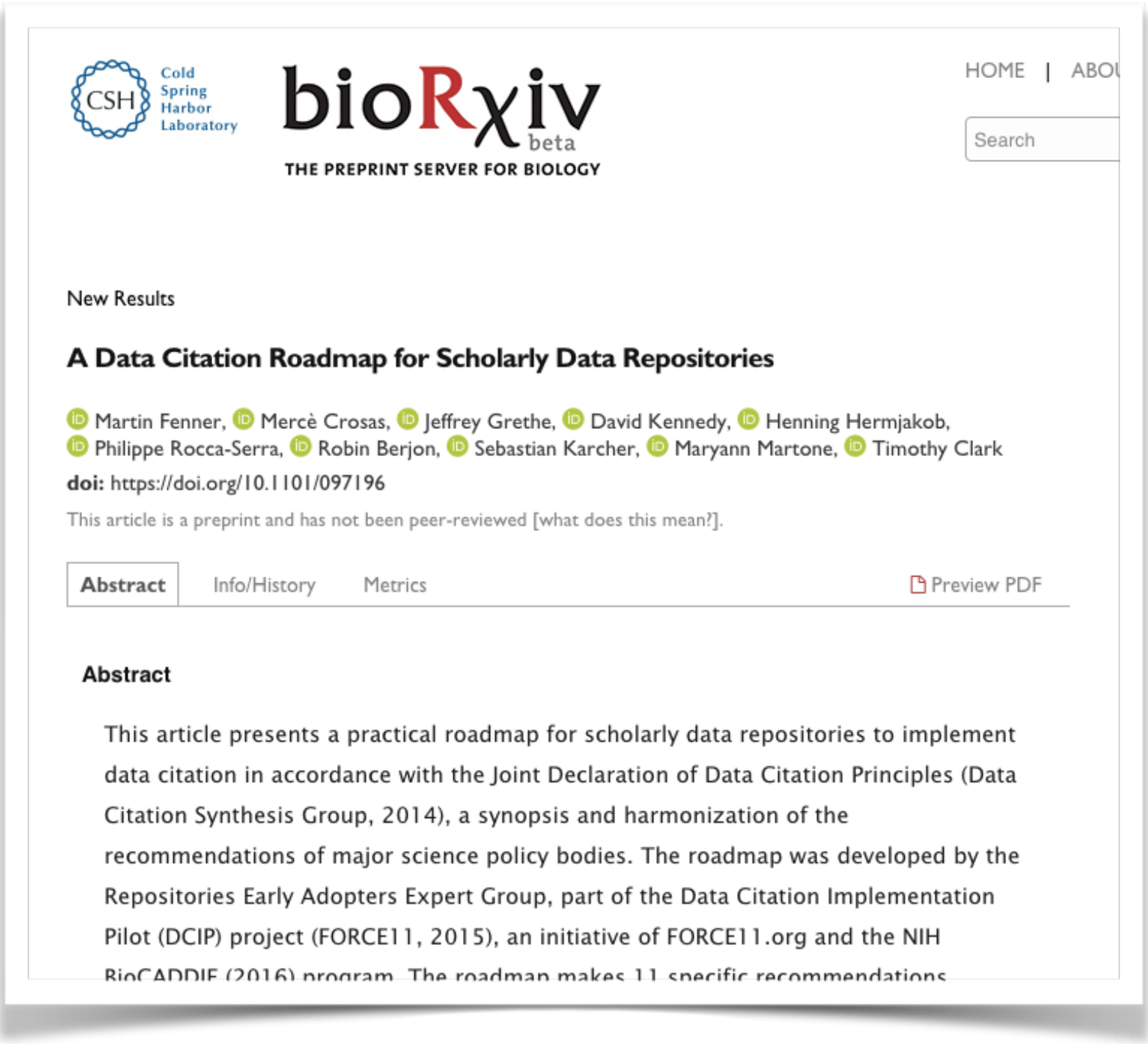
Altman M, Castro E, **Crosas M**, Durbin P, Garnett A, Whitney J. 2015, "Open Journal Systems and Dataverse Integration-- Helping Journals to Upgrade Data Publication for Reusable Research" Code4Lib Journal, Issue 30

Starr J, Castro E, **Crosas M**, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. 2015, "Achieving human and machine accessibility of cited data in scholarly publications" PeerJ Computer Science, 1:e1 https://dx.doi.org/10.7717/peerj-cs.1

Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer,K., **Crosas, M.,** Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D.W., Kashyap,V., Mahabal, A., Siemiginowska, A., Slavkovic, A., 2014. 10 Simple Rules for the Care and Feeding of Scientific Data, PLoS Comput Biol, doi:10.1371/journal.pcbi.1003542

Pepe, A., Goodman, A., Muench, A., **Crosas, M.,** Erdmann, C., 2014. How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE, DOI: 10.1371/journal.pone.0104798

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

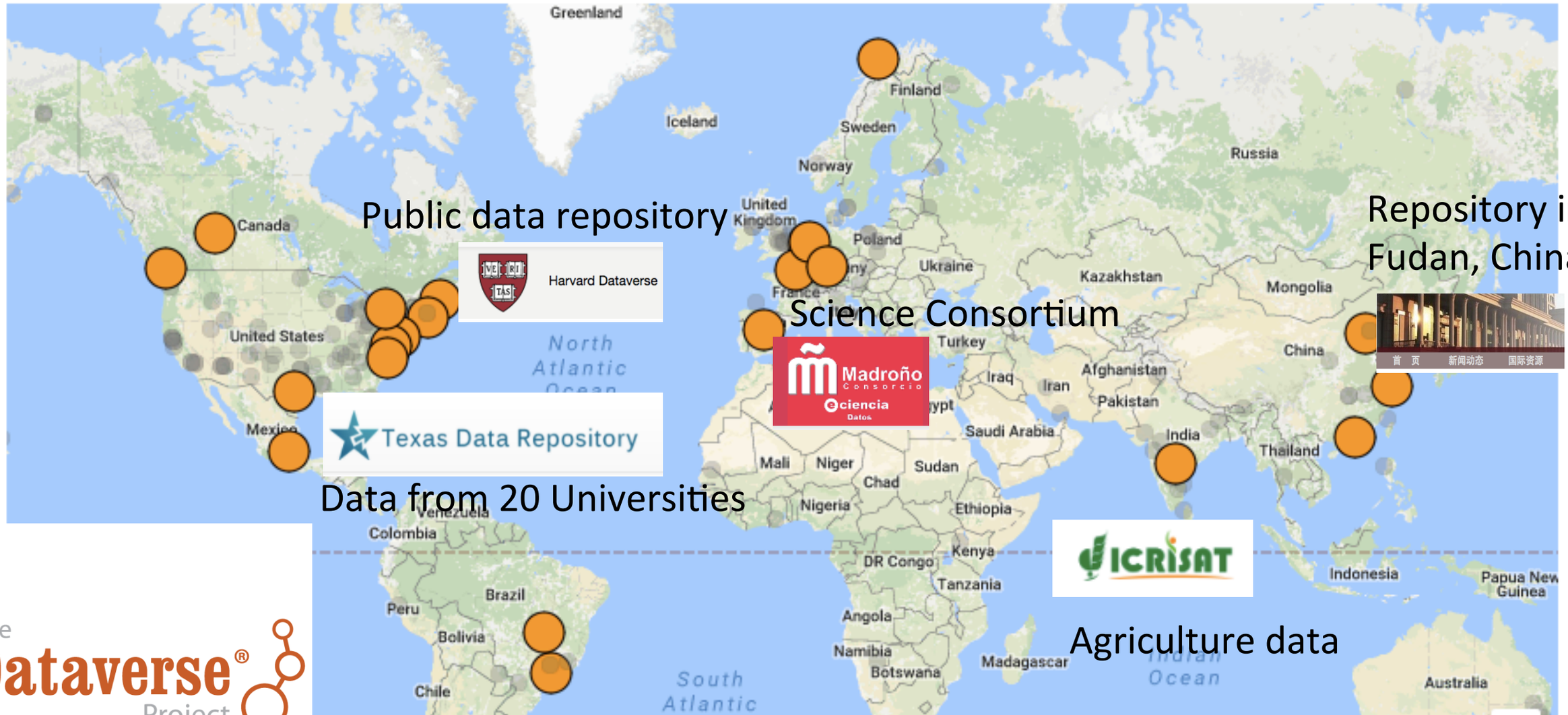
Fenner et al, 2016, "A Data Citation Roadmap for Scholarly Repositories" BioArxiv (preprint)
Synthesis Group, 2014, Joint Declaration of Data Citation Principles (Force11)

Dataverse

An open-source platform for building any type of data repository
With a growing community of developers and users

http://dataverse.org

22 Installations | 2,133 Dataverses | 48,690 Datasets | 2,400,322 Downloads



The Dataverse® Project

Created at IQSS at Harvard University

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

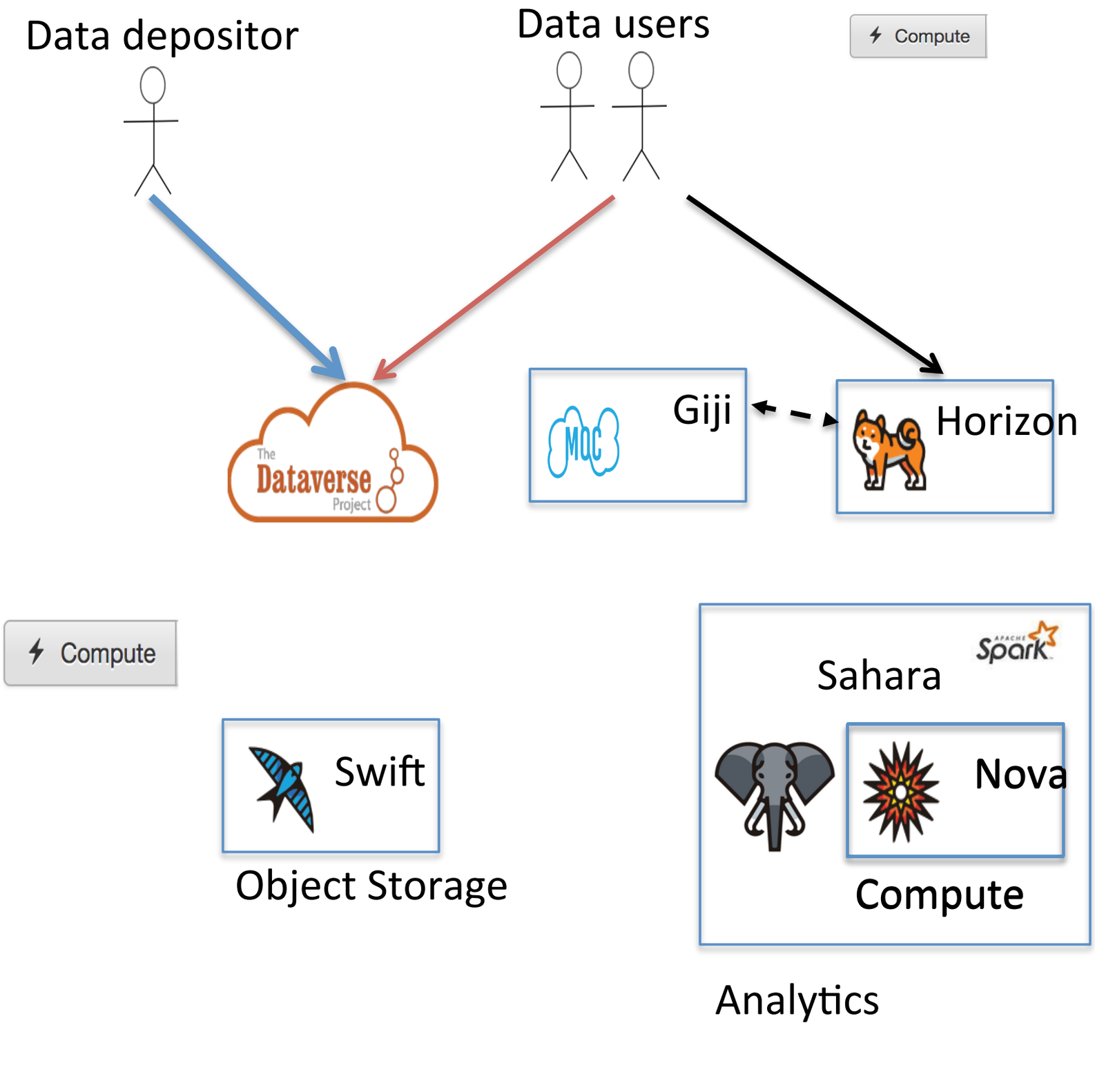
Current Challenges for Dataverse Repositories

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

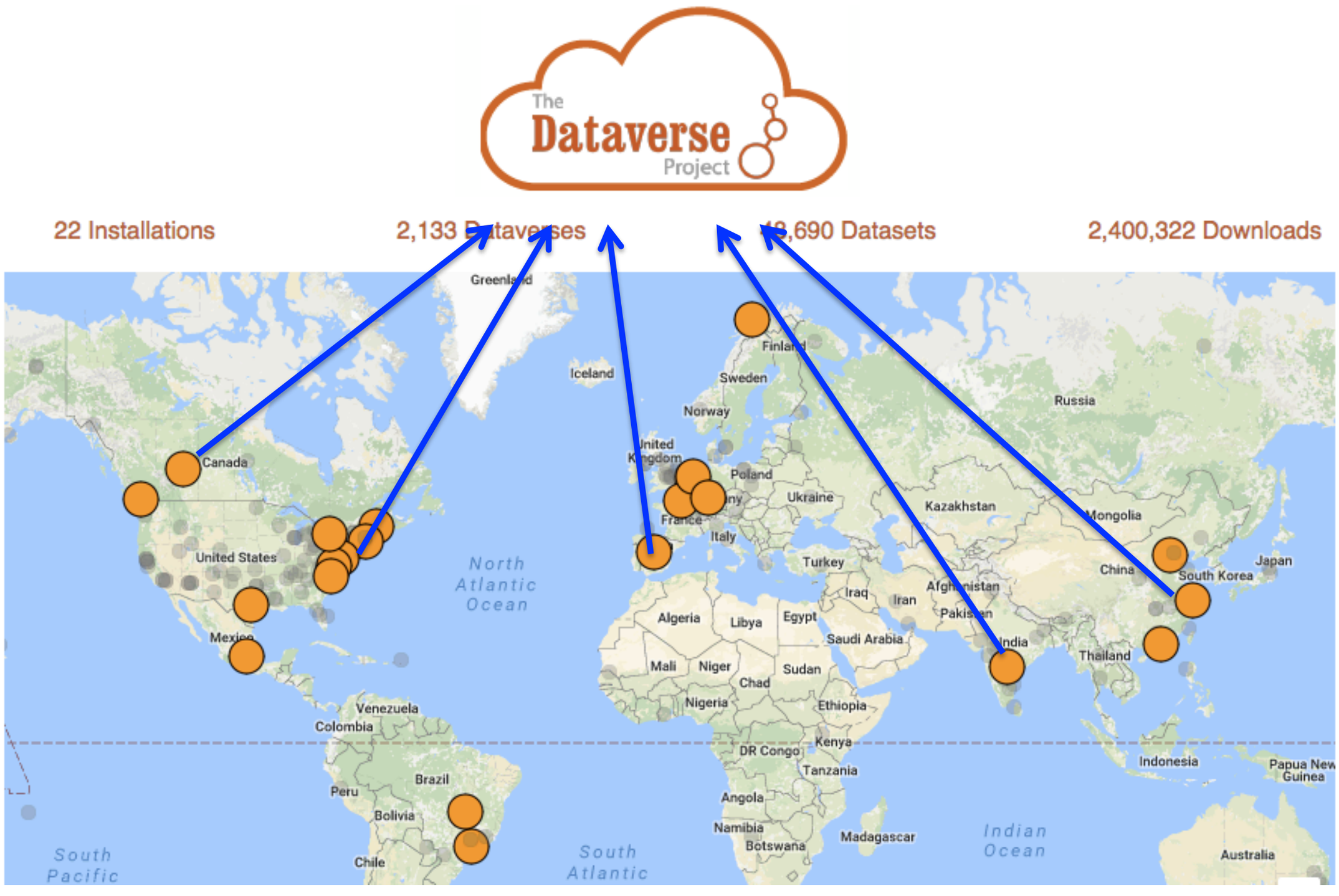
DATA REPOSITORIES NEED CLOUDS

CLOUDS NEED DATA REPOSITORIES

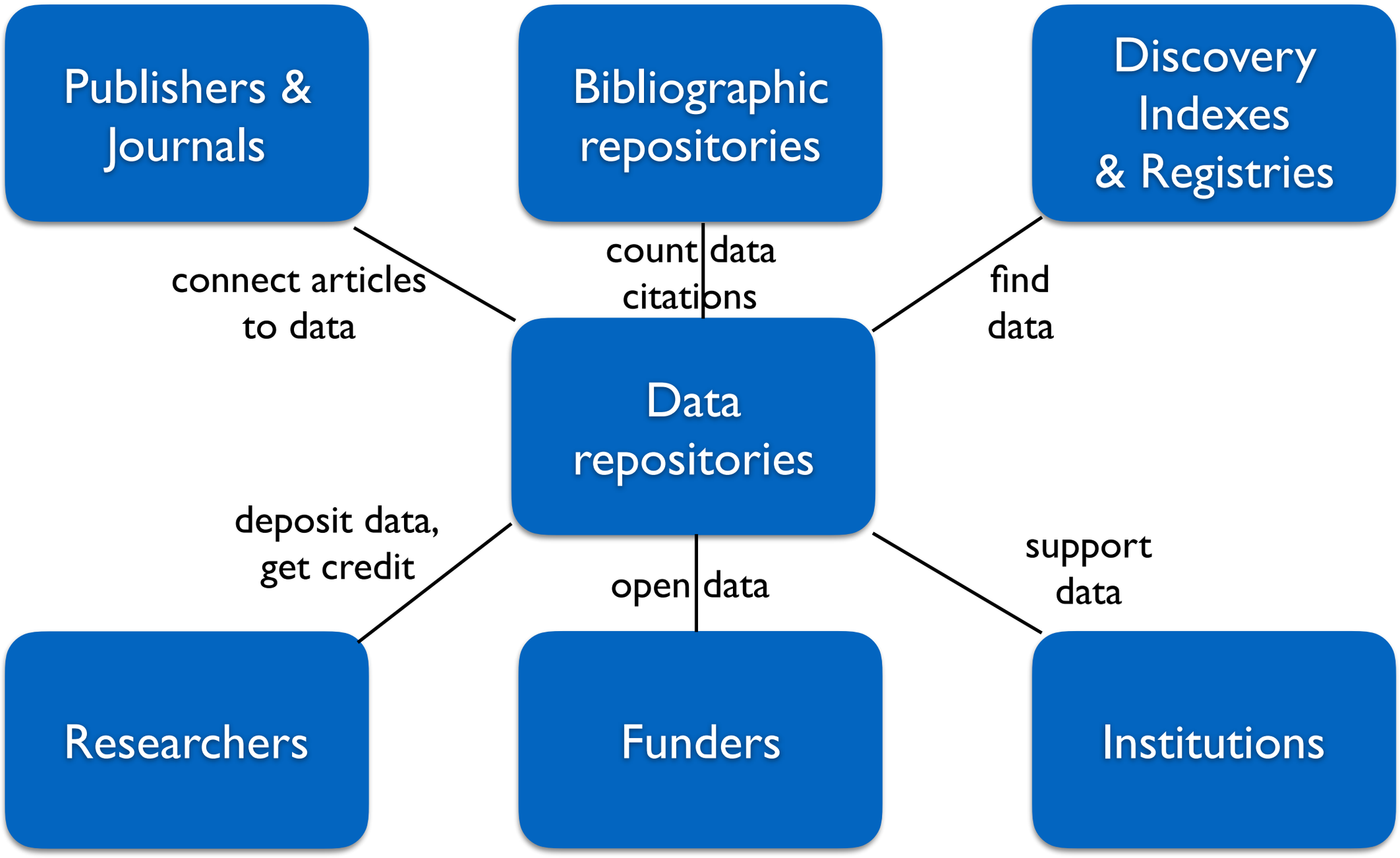
Cloud Dataverse on Massachusetts Open Cloud



World Wide Data Federation



In order to build data sharing incentives, all parties need to be on board



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