Building and Supporting Data Repositories

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

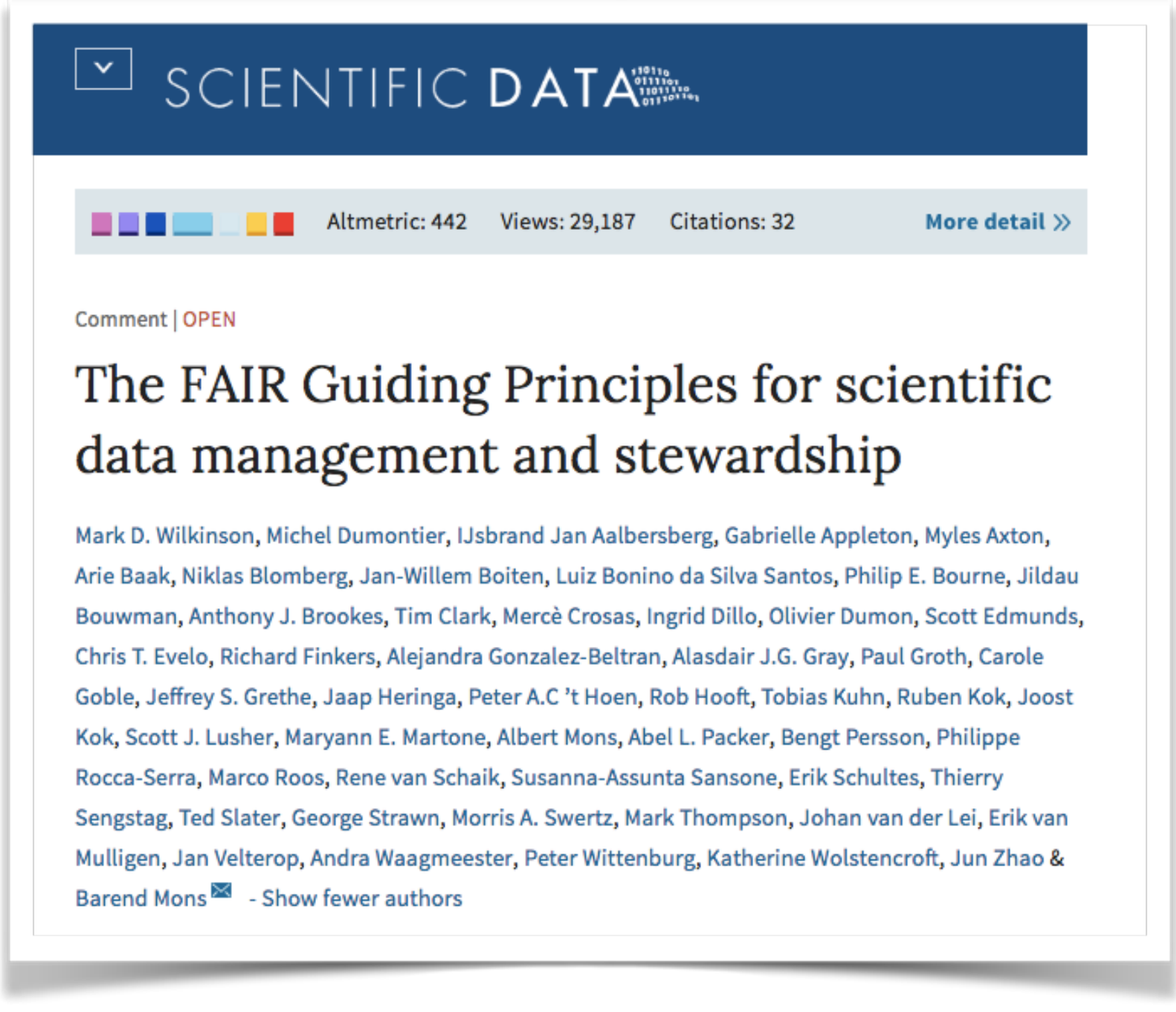
Qualitative Repositories Workshop, SESYNC, February 27, 2017

The importance of being FAIR

The importance of being cited

The importance of being connected

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;

"**FAIR Principles** put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals."

"**Good data management** is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process."

The importance of being FAIR

The importance of being cited

The importance of being connected

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

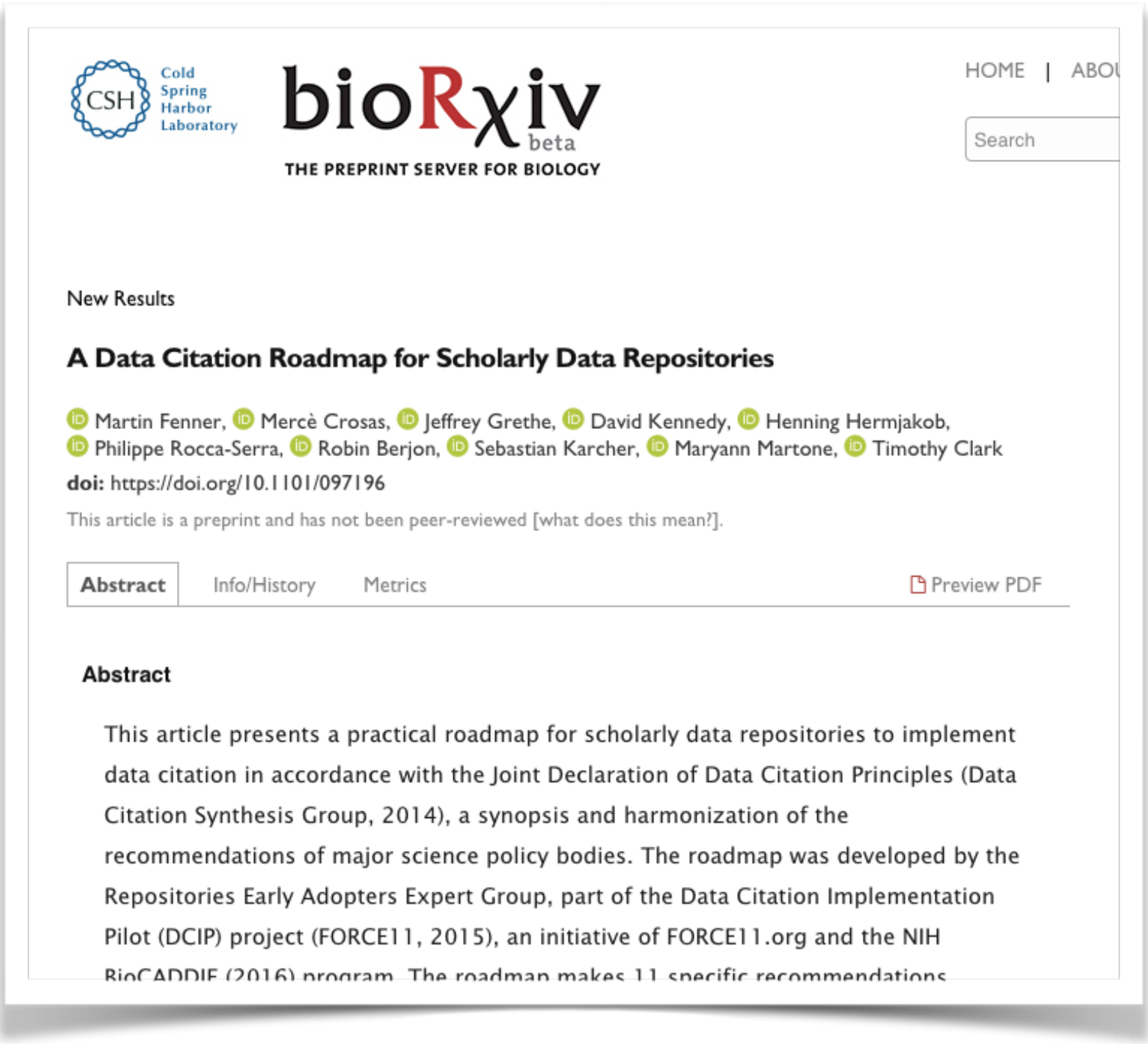
Altman M, Castro E, **Crosas M**, Durbin P, Garnett A, Whitney J. 2015, "Open Journal Systems and Dataverse Integration-- Helping Journals to Upgrade Data Publication for Reusable Research" Code4Lib Journal, Issue 30

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

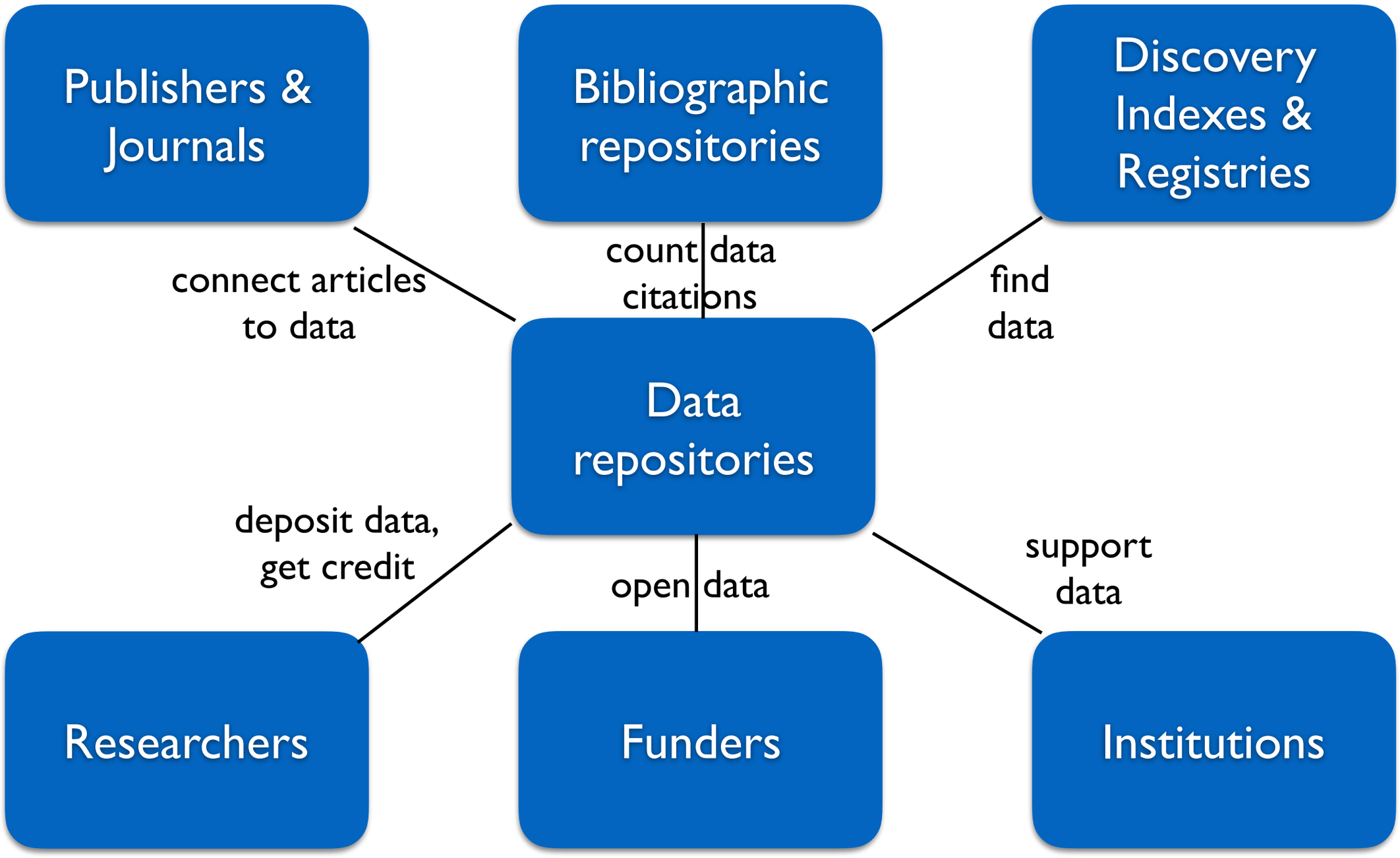
Fenner et al, 2016, "A Data Citation Roadmap for Scholarly Repositories" BioArxiv (preprint)
Synthesis Group, 2014, Joint Declaration of Data Citation Principles (Force11)

The importance of being FAIR

The importance of being cited

The importance of being connected

To build incentives and impact, all parties need to be on board



Dataverse

Also for Qualitative Data

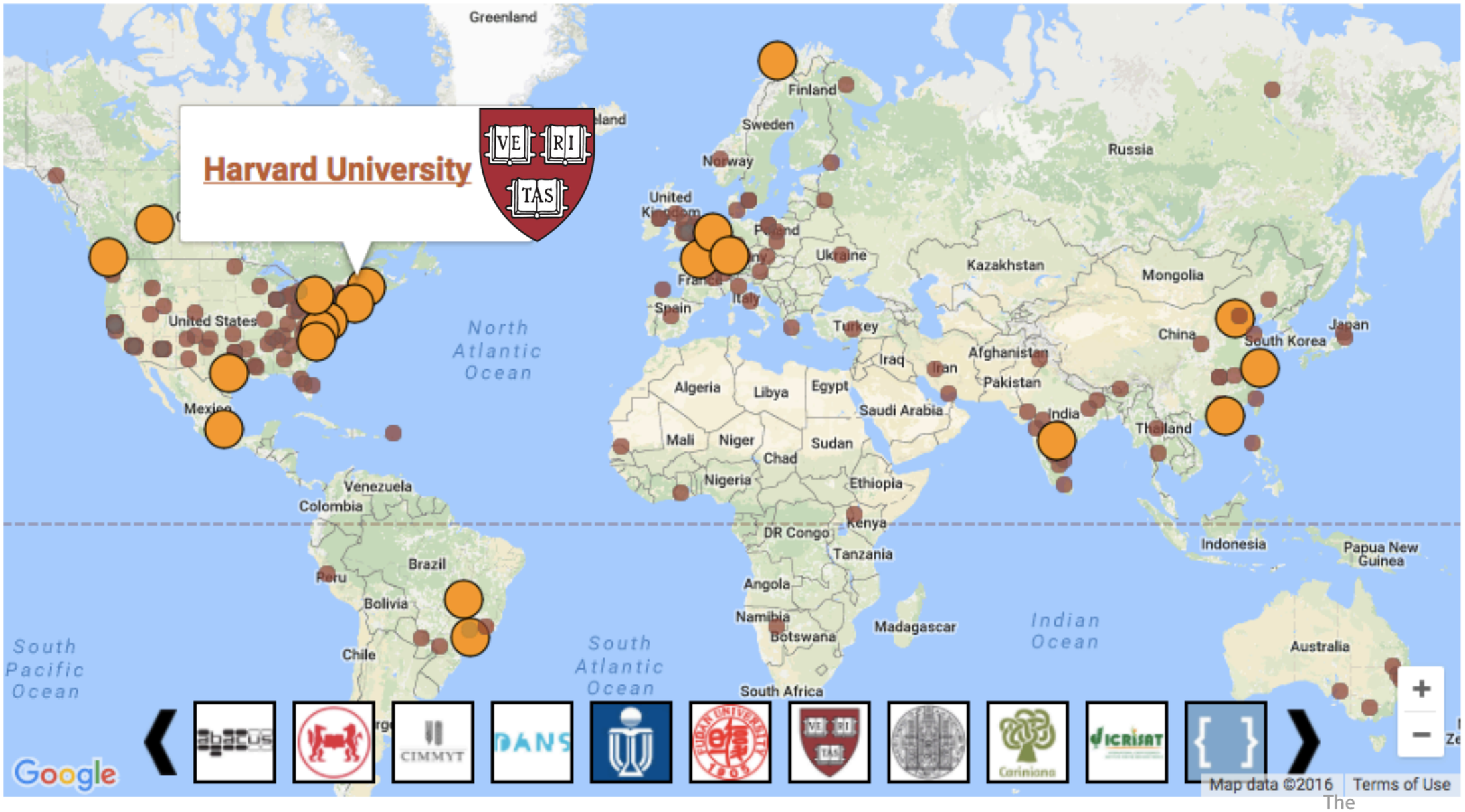
We built Dataverse to incentivize data sharing, with good data management in mind

- An **open-source** platform to share and archive data
- Developed at **Harvard's Institute for Quantitative Social Science (IQSS)** since 2006
- Gives **credit and control** to data authors & producers
- Implements **FAIR Principles** and **Data Citation** roadmap*
- Builds a **community** to:
  - define new standards and best practices
  - foster new research in data sharing and reproducibility
- Has brought **data publishing** into the hands of data authors

The Dataverse Project

Dataverse is now a widely used repository platform

22 installations around the world
70,000 datasets in Harvard Dataverse repository
Used by researchers from > 500 institutions
http://dataverse.org



The Dataverse Project

Dataset Landing Page

Dataverse

Metrics 271 Downloads

Data Citation, automatically registered to DataCite

Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse, V2

Cite Dataset

Learn about Data Citation Standards.

Description

Bike collisions in Boston (2009 - 2012). The data are derived from Boston Police Department (BPD) incident narrative reports, as organized and compiled by Dahianna Lopez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston Cyclists Union. Visitors can also explore the data through an interactive map hosted by BostonMap.

Files | Metadata | Terms | Versions

Search this dataset...

Files: data, docs, code
Any format accepted, some formats preferred

3 Files

Download

	Bike Collision Codebook.pdf Adobe PDF - 884.4 KB - Jun 8, 2014 - 82 Downloads MD5: b2fe76857594c44f67fef0102752cbeb Documentation	Download
	Bike Collisions.zip ZIP Archive - 256.9 KB - Jun 8, 2014 - 72 Downloads MD5: 4164ea35da5b8816908bc57ba8d72ef6 Shapefiles	Download
	Final Bike Collision Database.xlsx MS Excel (XLSX) - 1.7 MB - Jul 24, 2014 - 90 Downloads MD5: f3c25f076485f9c6791e3ef0e98e604f Raw Data	Download

Dataset Landing Page

Dataverse

About Guides Support Sign Up Log In

Metrics 271 Downloads

Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse, V2

Cite Dataset

Learn about Data Citation Standards.

Description

Bike collisions in Boston (2009 - 2012). The data are derived from Boston Police Department (BPD) incident narrative reports, as organized and compiled by Dahianna Lopez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston Cyclists Union. Visitors can also [illegible] map hosted by BostonMap.

Basic Citation metadata required, rich metadata recommended

Files Metadata Terms Versions

Export Metadata

Citation Metadata

Dataset Persistent ID	doi:10.7910/DVN/24[illegible]
Publication Date	2014-06-09
Title	Bicycle Collisions in Boston, MA (2009-2012)
Author	Lopez, Dahianna (Harvard University) Cheevers, Maria (Boston Police Department) Stidman, Pete (Boston Cyclists Union) O'Brien, Daniel (Boston Area Research Initiative)
Description	Bike collisions in Boston (2009 - 2012). The data are derived from Boston Police Department (BPD) incident narrative reports, as organized and compiled by Dahianna Lopez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston Cyclists Union. Visitors can also explore the data through an interactive map hosted by BostonMap.
Producer	Boston Area Research Initiative (Harvard University)

Dataset Landing Page

Dataverse

About Guides Support Sign Up Log In

Metrics 271 Downloads

Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse, V2

Cite Dataset

Learn about Data Citation Standards.

Description

Bike collisions in Boston (2
organized and compile
Cyclists Union. Visito
reports, as
Boston

Terms of Use & Licenses: open data encouraged, restrictions optional; DataTags to enable sharing sensitive data (coming soon)

Files Metadata Terms Versions

Terms of Use

Waiver

Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse.

CC0 - "Public Domain Dedication"

Guestbook

Guestbook

The following guestbook will prompt a user to provide additional information when downloading a file.

Boston Area Research Initiative Preview Guestbook

Dataset Landing Page

Bicycle Collisions in Boston, MA (2009-2012) Dataverse (Harvard University, Northeastern University)

Metrics 271 Downloads

Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse, V2

Cite Dataset

Learn about Data Citation Standards.

Description

Bike collisions in Boston (20 ... Boston Police Department (BPD) incident narrative reports, as organized and com... arch Center) and partners at BPD, BARI and the Boston Cyclists Union... hosted by BostonMap.

Datasets and file versions, with provenance (coming soon)

Files | Metadata | Terms | Versions

View Differences

☐	2.1	View Details	Dan O'Brien	February 4, 2016
☐	2.0	**Files (Added: 1; Removed: 1);** View Details	Dan O'Brien	July 25, 2014
☐	1.0	This is the first published version.	Dahianna Lopez, Dan O'Brien	June 6, 2014

Thanks!

Learn more at http//dataverse.org

@mercecrosas mercecrosas.com