

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

Measuring the Impact of Digital Repositories, March 1, 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

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:jNUA7tB3bFeMcC2oGBvdlHw==

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



The screenshot shows the bioRxiv preprint server interface. At the top left is the Cold Spring Harbor Laboratory (CSH) logo. The bioRxiv logo is prominently displayed in the center, with the tagline 'THE PREPRINT SERVER FOR BIOLOGY'. A search bar is located in the top right corner. Below the header, the article title 'A Data Citation Roadmap for Scholarly Data Repositories' is shown, followed by the authors' names: Martin Fenner, Mercè Crosas, Jeffrey Grethe, David Kennedy, Henning Hermjakob, Philippe Rocca-Serra, Robin Berjon, Sebastian Karcher, Maryann Martone, and Timothy Clark. The DOI is provided as https://doi.org/10.1101/097196. A note indicates that the article is a preprint and has not been peer-reviewed. Below the article information, there are tabs for 'Abstract', 'Info/History', and 'Metrics', along with a 'Preview PDF' button. The abstract text is visible below the tabs.

Required:

- persistent ID/url resolves to dataset landing page

Recommended:

- landing page includes human- and machine-readable metadata

Optional:

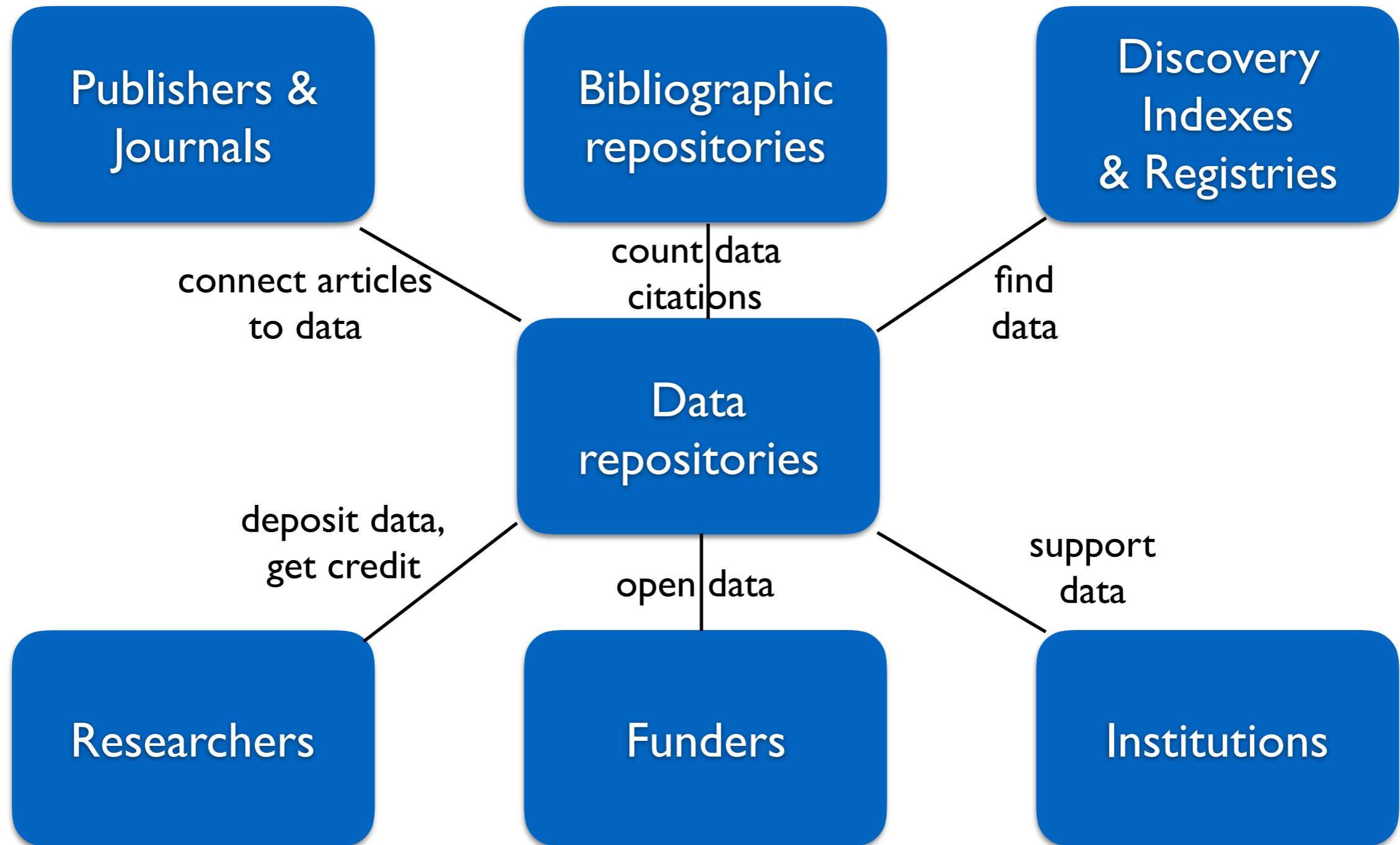
- content negotiation for more accessible metadata

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: a FAIR repository software with a growing community

22 installations around the world; growing active community of developers and users

Harvard Dataverse repository: 73,000 datasets, 330,000 files

23,000 datasets deposited at Harvard by researchers from > 500 institutions

A cost-effective way to build a new data repository

<http://dataverse.org>

