FAIR Data Management
FAIR Data Sharing

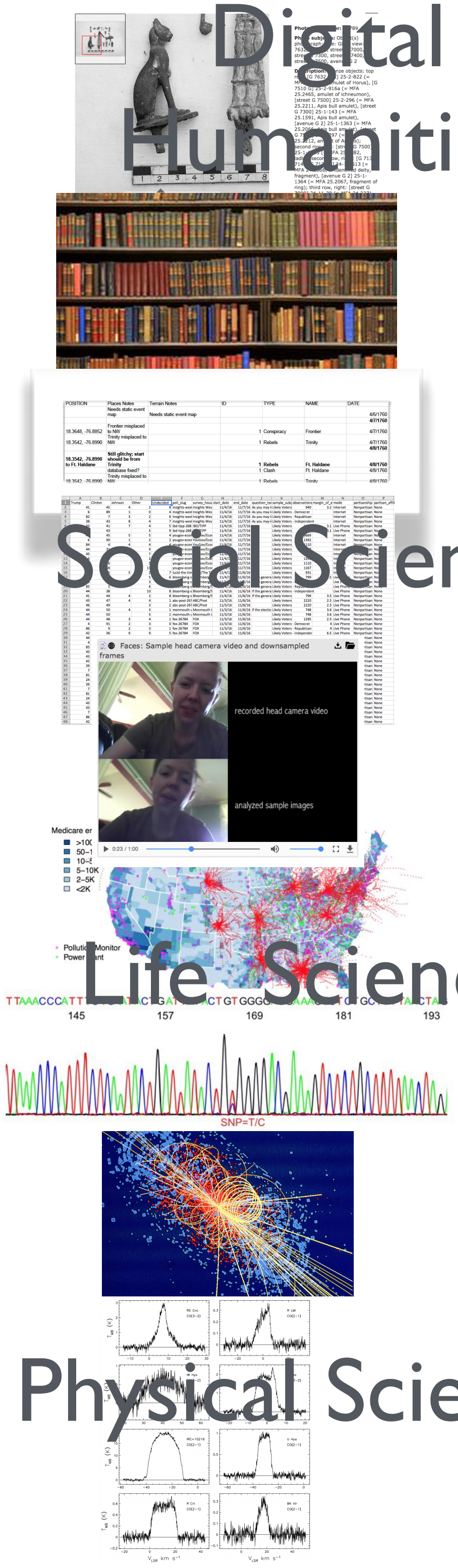
Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University
@mercecrosas mercecrosas.com

Critical Perspectives of on the Practice of Digital Archeology, Harvard University, February 3, 2017

primary data

order, transform, & analyze them

gain knowledge, make decisions



Digital Humanities

Social Sciences

Life Sciences

Physical Sciences

Catalog
Classify
Visualize
Quantify
Summarize
Geo reference
Inference
Missing data
Forecast
Causal Inference
Coding
Annotations
Associations
Likelihoods
Compare with theory

Learn about the whole from a part.

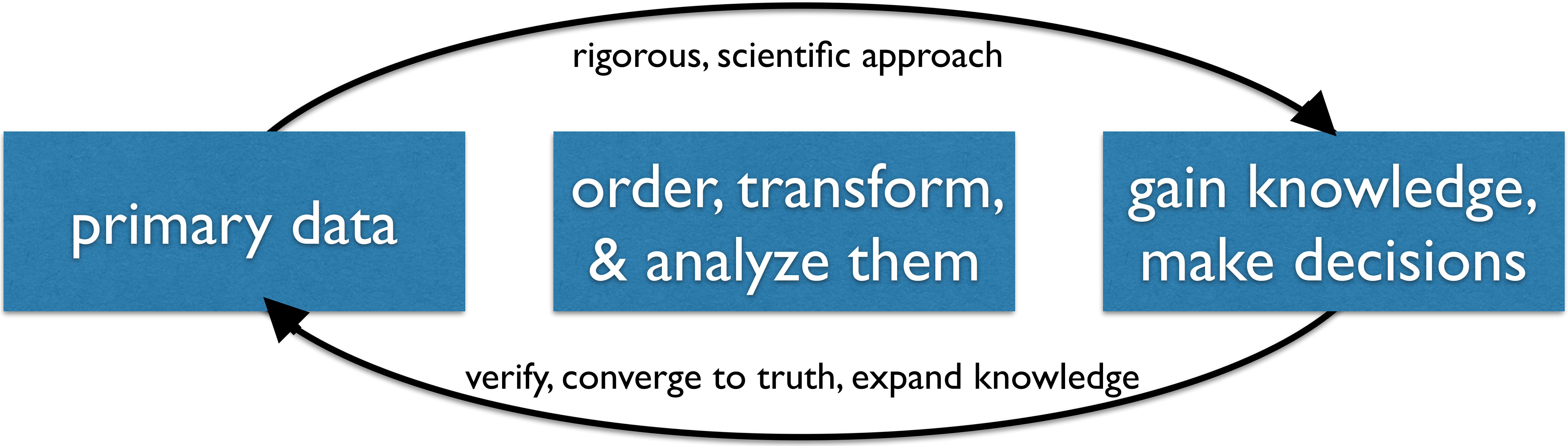
Tell a story.

Make a prediction.

Ultimately explain.

Nullius in Verba: “Take nobody’s word for it”

(Royal Society, Philosophical Transactions, 1965)



rigorous, scientific approach

primary data

order, transform, & analyze them

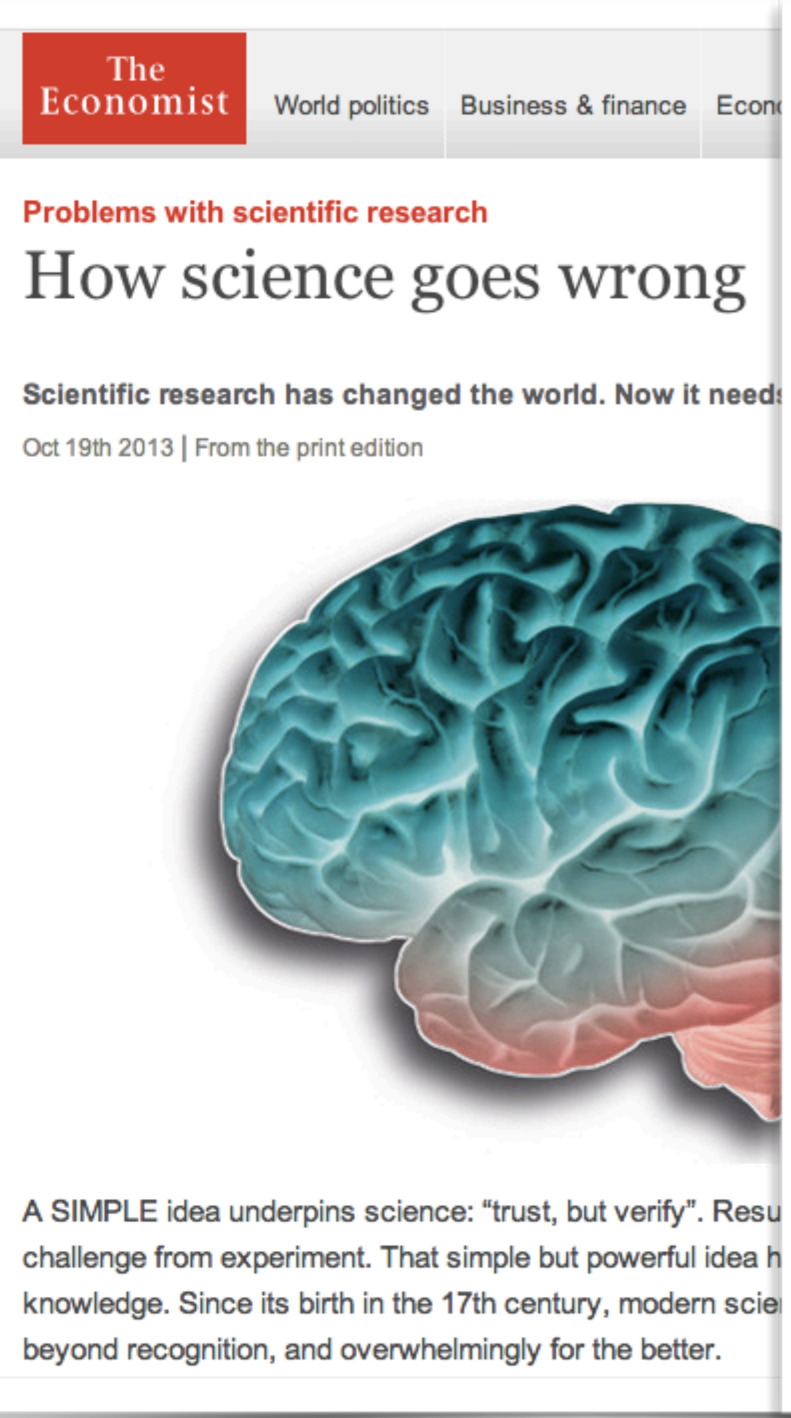
gain knowledge, make decisions

verify, converge to truth, expand knowledge

- **Replication:** Independent scientific experiments to validate findings
- **Reproducibility:** Calculation of quantitative results by others using original datasets and methods

(Definitions by Stodden, Leisch, Peng, Implementing Reproducible Research, 2014)
* Replication and reproducibility definitions vary across disciplines

“Answering even a simple scientific question requires lots of choices that can shape the results”



The Economist

World politics Business & finance Econ

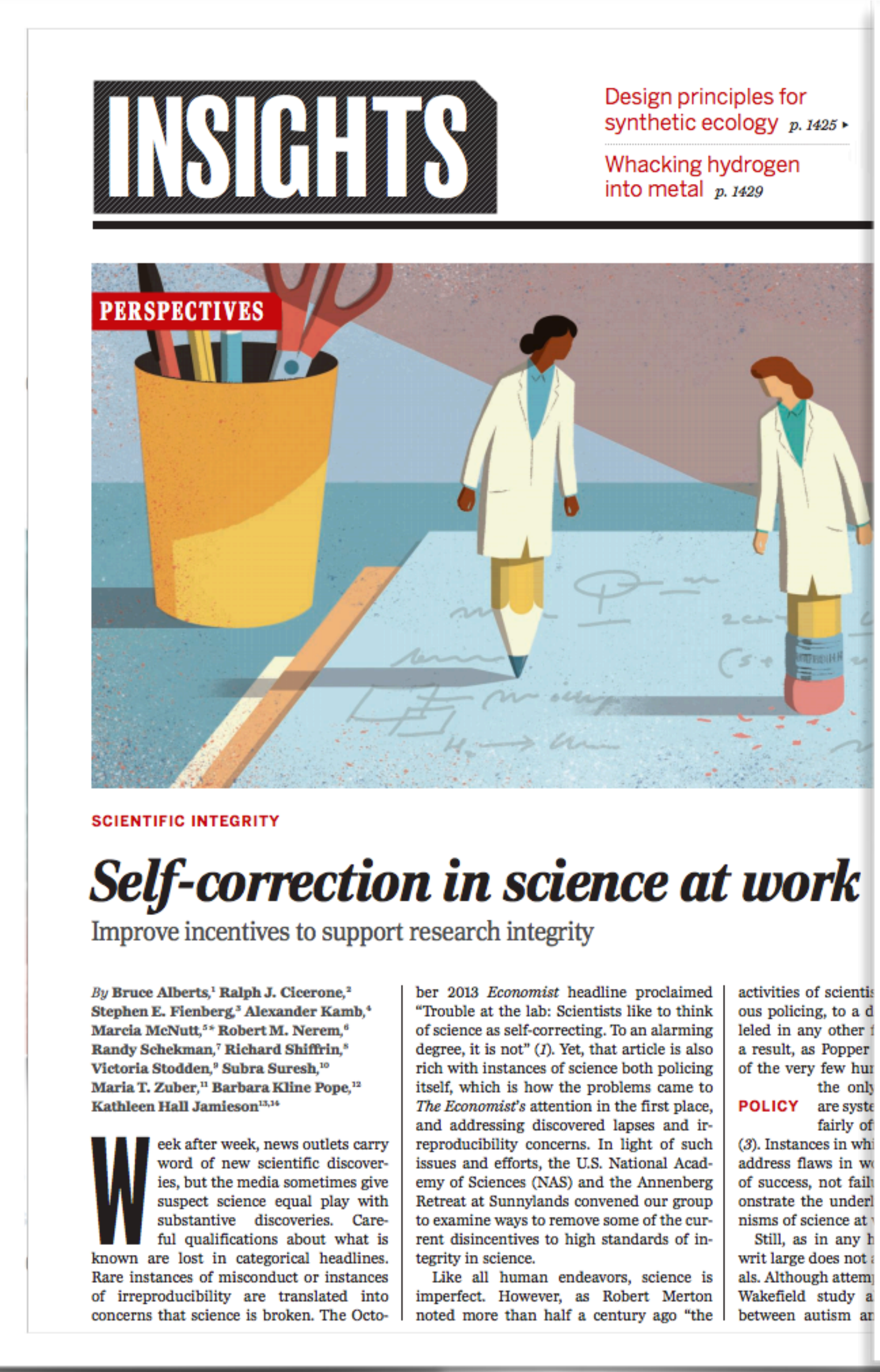
Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs

Oct 19th 2013 | From the print edition

A SIMPLE idea underpins science: “trust, but verify”. Resu
challenge from experiment. That simple but powerful idea h
knowledge. Since its birth in the 17th century, modern scie
beyond recognition, and overwhelmingly for the better.



INSIGHTS

Design principles for synthetic ecology *p. 1425* ▸

Whacking hydrogen into metal *p. 1429*

PERSPECTIVES

SCIENTIFIC INTEGRITY

Self-correction in science at work

Improve incentives to support research integrity

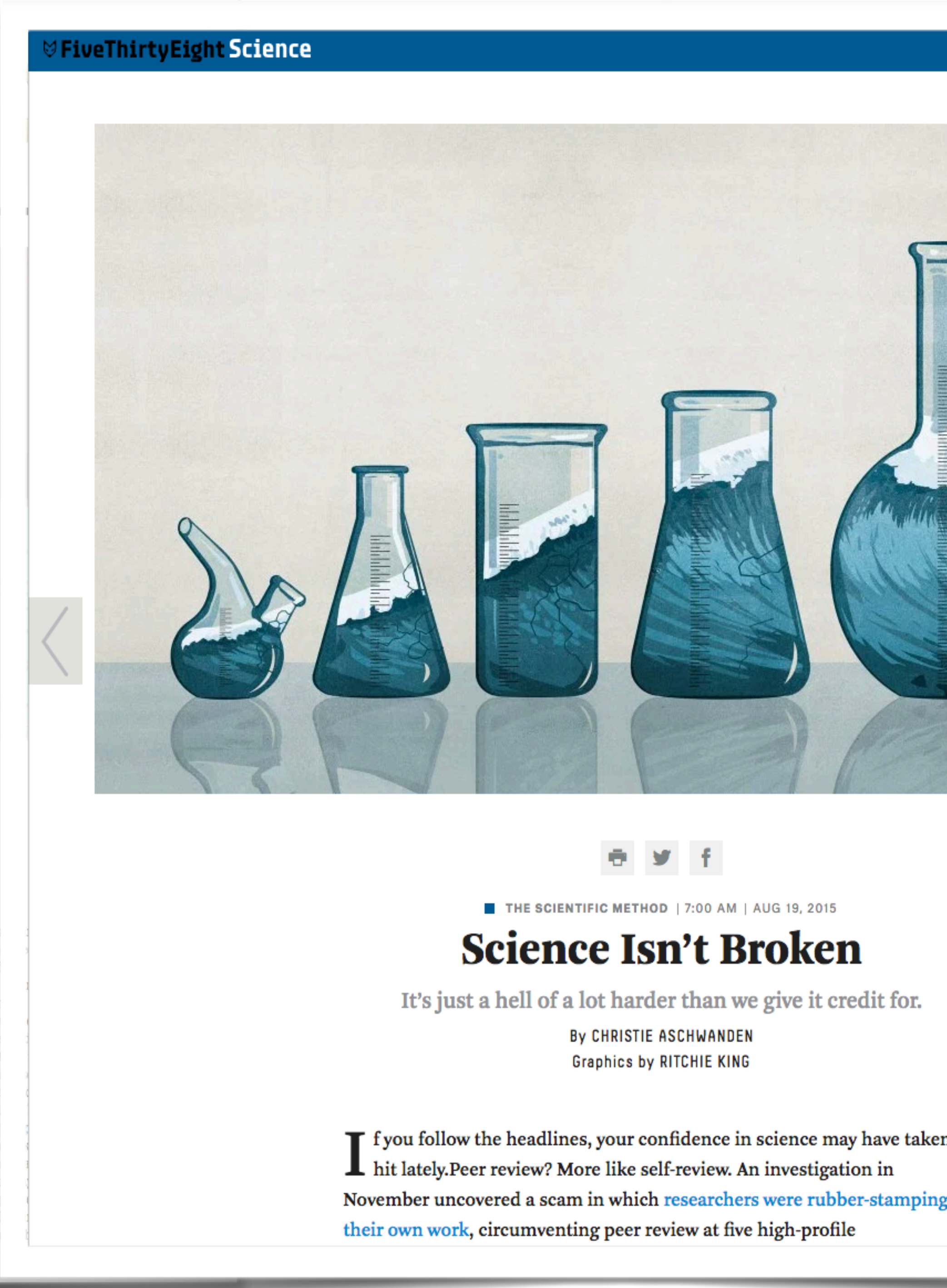
By **Bruce Alberts,[1] Ralph J. Cicerone,[2] Stephen E. Fienberg,[3] Alexander Kamb,[4] Marcia McNutt,[5]* Robert M. Nerem,[6] Randy Schekman,[7] Richard Shiffrin,[8] Victoria Stodden,[9] Subra Suresh,[10] Maria T. Zuber,[11] Barbara Kline Pope,[12] Kathleen Hall Jamieson[13,14]**

Week after week, news outlets carry word of new scientific discoveries, but the media sometimes give suspect 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 Annenberg Retreat at Sunnylands convened our group 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 scienti
ous policing, to a d
leled in any other
a result, as Popper
of the very few hu
the only
are syste
fairly of
(*3*). Instances in wh
address flaws in w
of success, not fail
onstrate the underl
nisms of science at

POLICY

Still, as in any h
writ large does not
als. Although attem
Wakefield study a
between autism a



FiveThirtyEight Science

THE SCIENTIFIC METHOD | 7:00 AM | AUG 19, 2015

Science Isn’t Broken

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

By CHRISTIE ASCHWANDEN
Graphics by RITCHIE KING

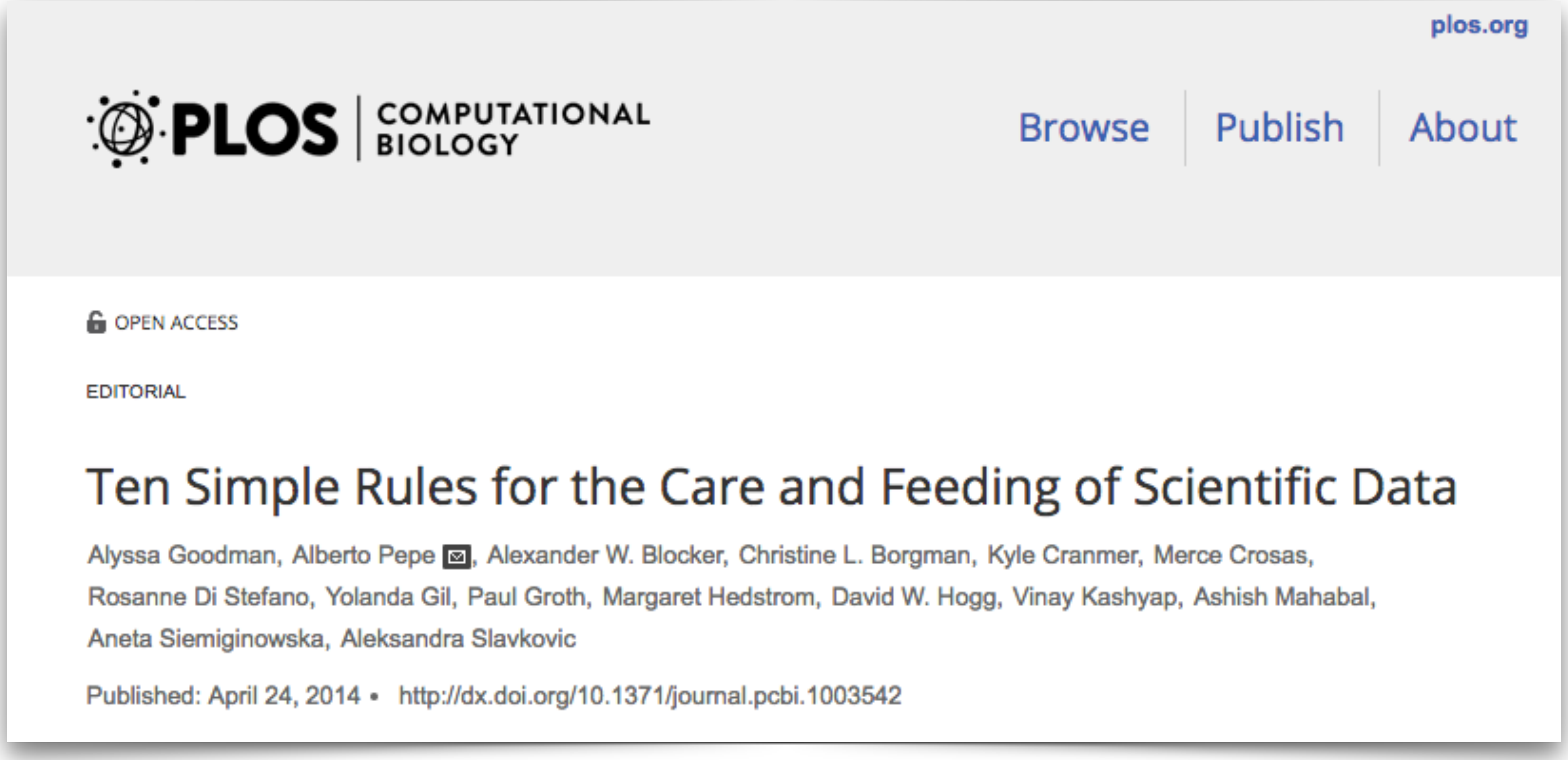
If you follow the headlines, your confidence in science may have taken
hit lately.Peer review? More like self-review. An investigation in
November uncovered a scam in which researchers were rubber-stamping
their own work, circumventing peer review at five high-profile

“When possible, make data, methods, and code open to verify”

“Science/research might be imperfect, but is self-correcting”

“It’s not unreliable, but more challenging that we give it credit for”

Caring for and sharing your data (and code) enable you and others to correct and reuse them



plos.org

PLOS | COMPUTATIONAL BIOLOGY

Browse | Publish | About

OPEN ACCESS

EDITORIAL

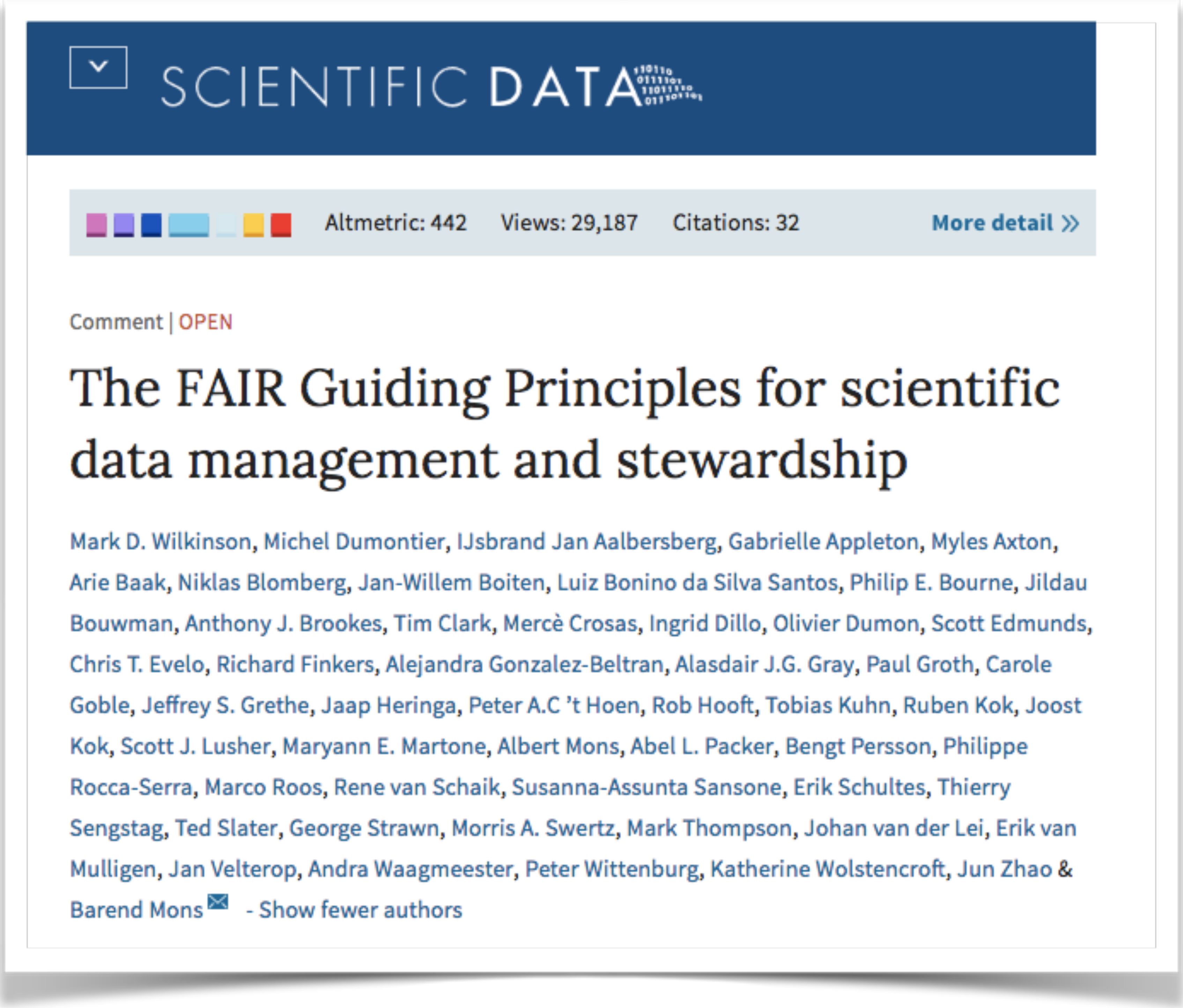
Ten Simple Rules for the Care and Feeding of Scientific Data

Alyssa Goodman, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Merce Crosas, Rosanne Di Stefano, Yolanda Gil, Paul Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, Aleksandra Slavkovic

Published: April 24, 2014 • http://dx.doi.org/10.1371/journal.pcbi.1003542

1. Love your data 2. Share your data 3. Conduct science with reuse in mind 4. Publish workflow 5. Link data to publications 6. Publish your code 7. State how you want to get credit 8. Foster and use repositories 9. Reward colleagues who share 10. Boost Data Science

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

Wilkinson et al, ‘The FAIR Guiding Principles scientific data management and stewardship,” Nature Scientific Data, 2016;
NIH Data Commons Principles; Joint Declaration of Data Citation Principles (Force11)

“**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.”

FAIR Data Principles in Brief

- **To be Findable:**
  - (meta)data are assigned a globally unique and persistent identifier
  - data are described with rich metadata
  - metadata clearly and explicitly include the identifier of the data it describes
  - (meta)data are registered or indexed in a searchable resource
- **To be Accessible:**
  - (meta)data are retrievable by their identifier using a standardized communications protocol
  - the protocol is open, free, and universally implementable
  - the protocol allows for an authentication and authorization procedure, where necessary
  - metadata are accessible, even when the data are no longer available
- **To be Interoperable:**
  - (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - (meta)data use vocabularies that follow FAIR principles
  - (meta)data include qualified references to other (meta)data
- **To be Reusable:**
  - meta(data) are richly described with a plurality of accurate and relevant attributes
  - (meta)data are released with a clear and accessible data usage license (meta)data are associated with detailed provenance (meta)data meet domain-relevant community standards

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 since 2006
- Gives **credit and control** to researchers
- Builds a **community** to:
  - define new standards and best practices
  - foster new research and collaboration in data sharing
- Has brought **data publishing** into the hands of researchers

The Dataverse Project

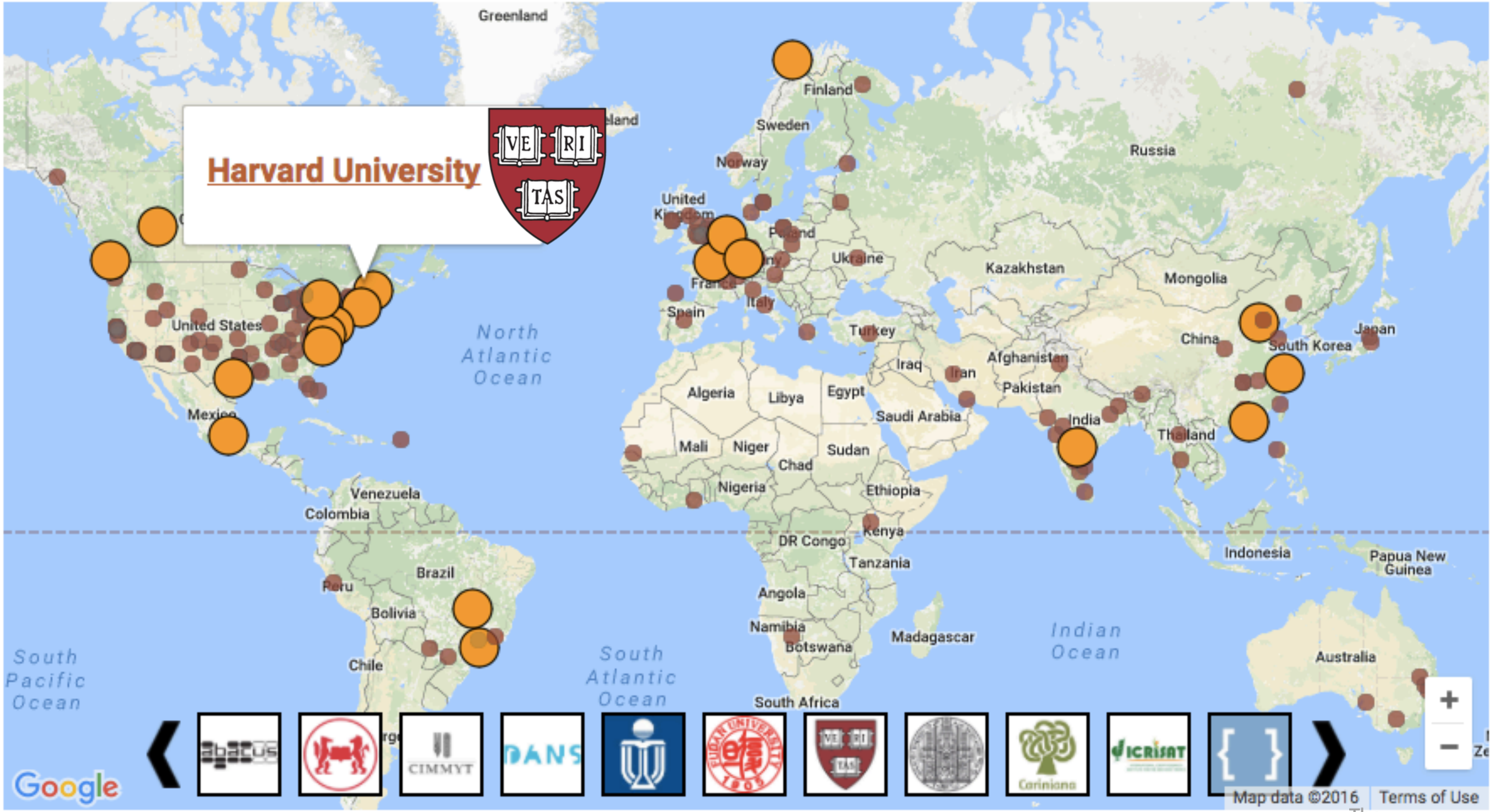
Dataverse is now a widely used repository platform

21 installations around the world

Used by researchers from > 500 institutions

60,000 datasets in Harvard Dataverse repository

http://dataverse.org



The Dataverse Project

Dataverse has a growing, engaged community of developers and users

38

GitHub contributors

332

members in the community list

23

community calls

with 239

participants from

8 countries

Annual Community Meeting,

with 200 attendees

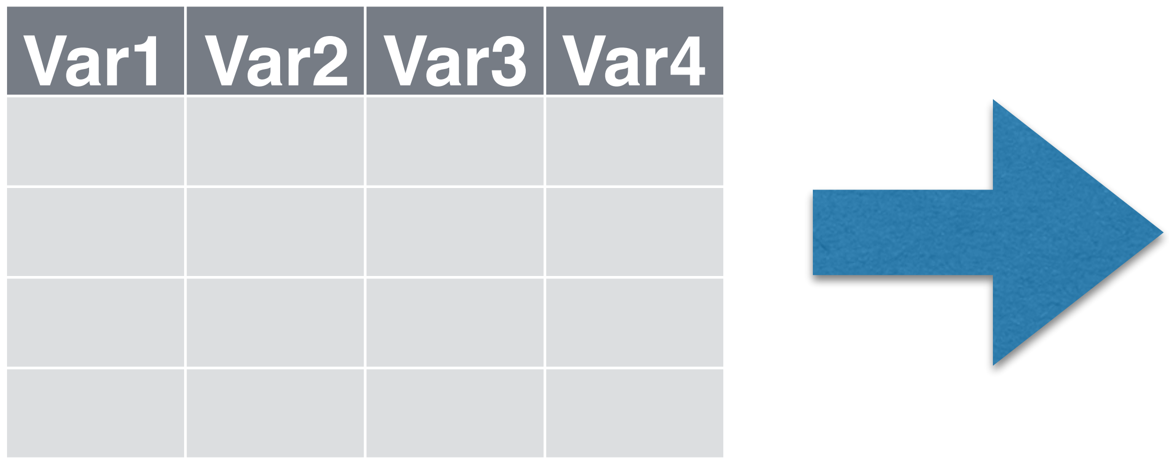
The Dataverse Project

Dataverse implements FAIR Data Principles

- **Data Citation with global persistent IDs:**
  - Generate DOI automatically
  - attribution to data authors and repository
  - registration to DataCite
- **Rich Metadata:**
  - citation metadata
  - domain-specific descriptive metadata
  - variable and file metadata (extracted automatically)
- **Access and usage controls:**
  - open data as default, with CC0 waiver
  - custom terms of use and licenses, when needed
  - data can be restricted, but citation & metadata always publicly accessible
- **APIs and standards:**
  - SWORD, OAI-PMH, Dataverse native open API
  - Dublin Core and DDI metadata standards
  - PROV ontology standard to capture provenance of a dataset (coming soon)

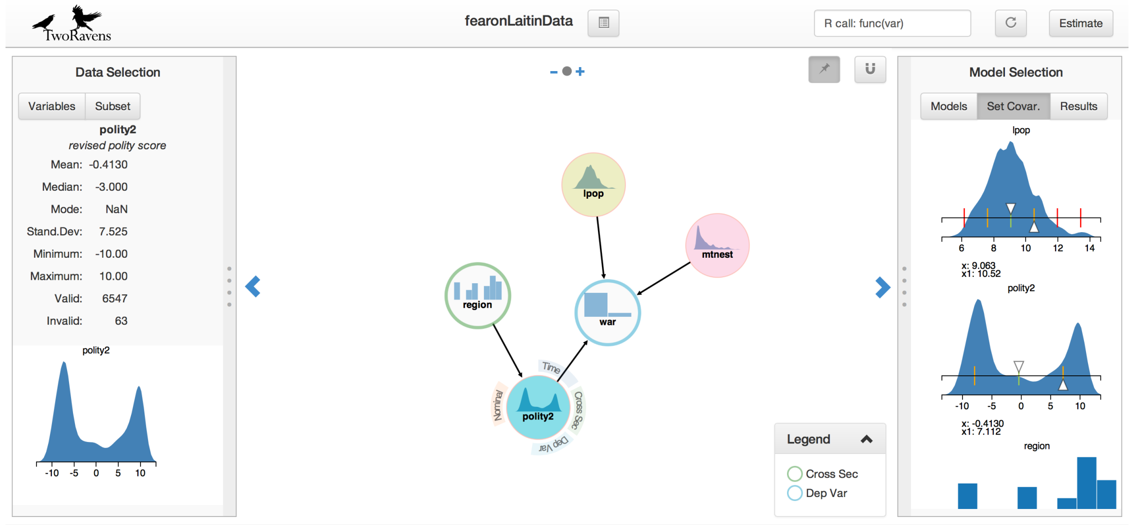
The Dataverse Project

Standard file formats and automatic metadata extraction allow data exploration



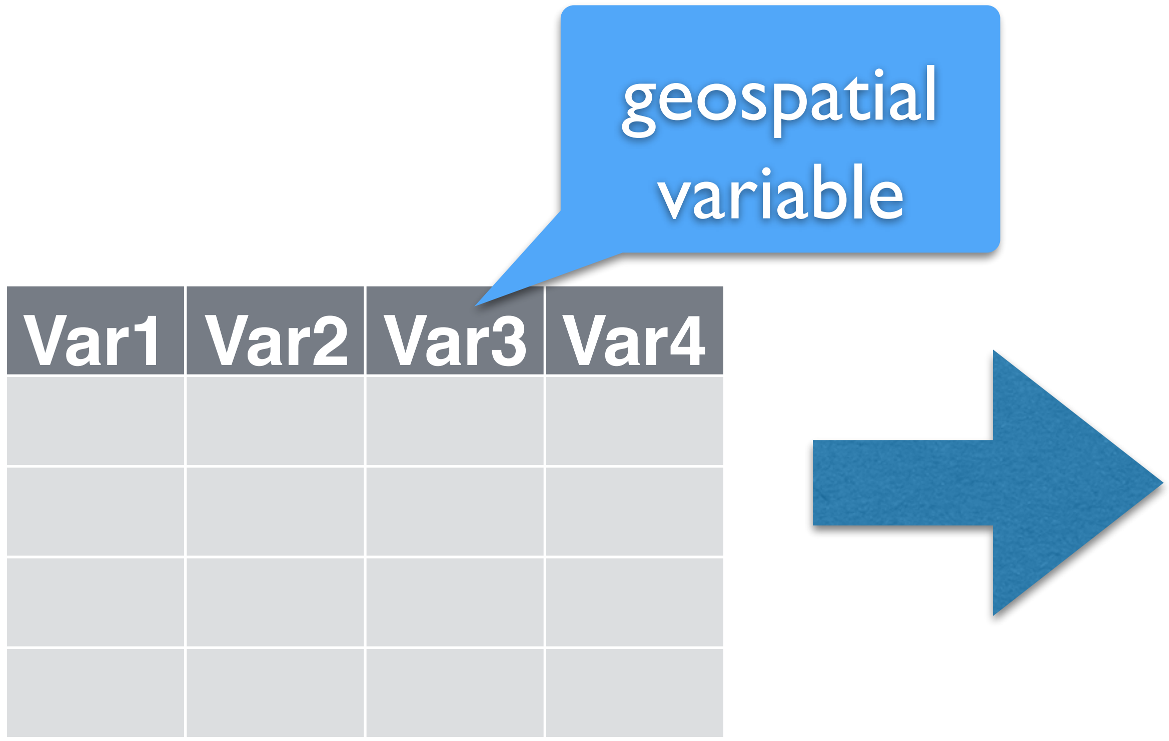
Var1	Var2	Var3	Var4

TwoRavens: summary stats & analysis



polity2
revised polity score

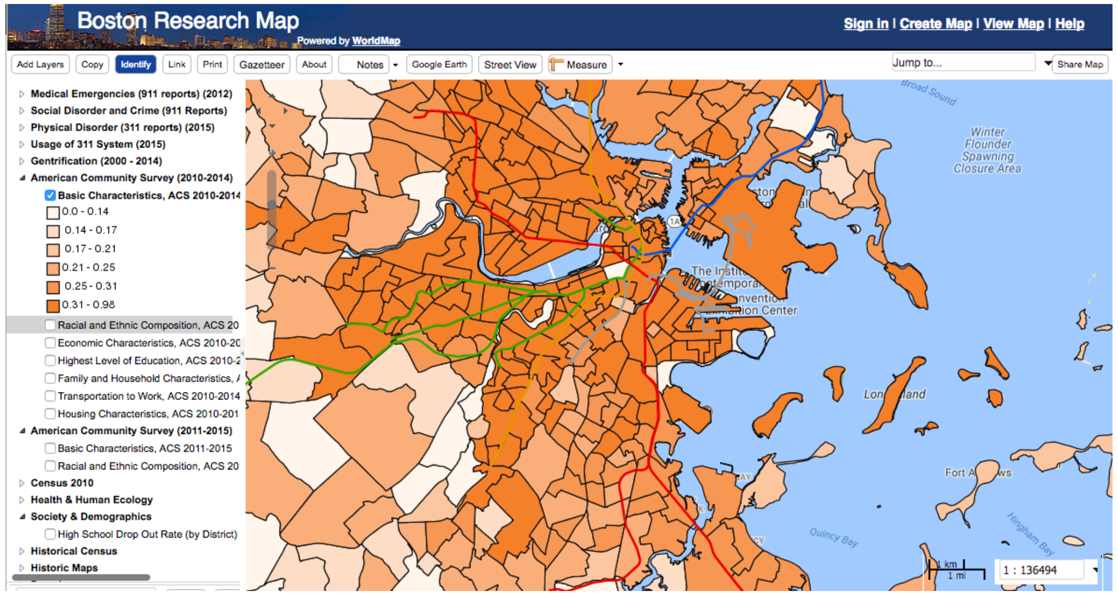
Mean:	-0.4130
Median:	-3.000
Mode:	NaN
Stand.Dev:	7.525
Minimum:	-10.00
Maximum:	10.00
Valid:	6547
Invalid:	63



geospatial variable

Var1	Var2	Var3	Var4

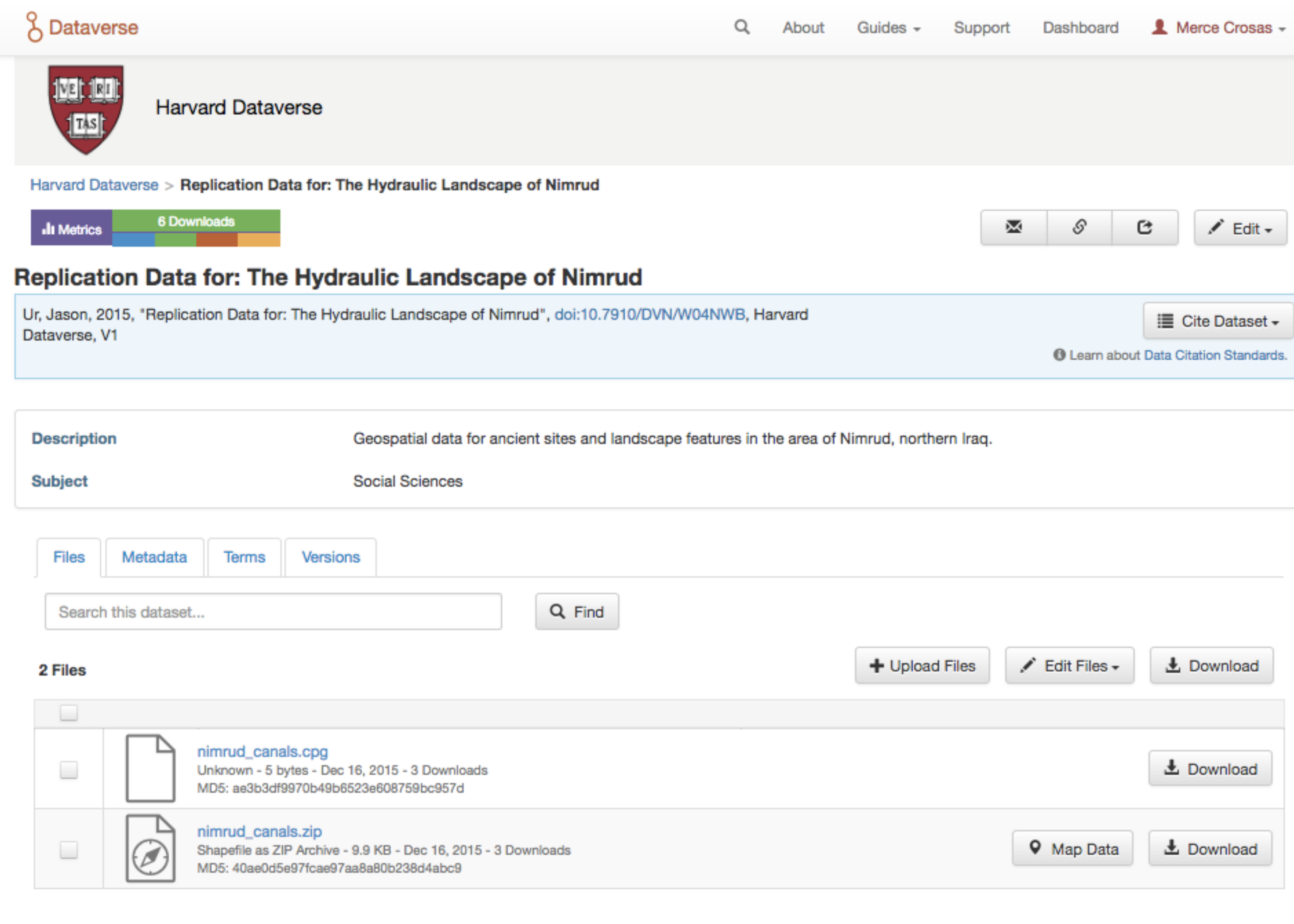
WorldMap: geospatial exploration



Boston Research Map
Powered by WorldMap
Sign in | Create Map | View Map | Help
Add Layers
Copy
Identify
Link
Print
Gazetteer
About
Notes
Google Earth
Street View
Measure
Jump to...
Share Map
Medical Emergencies (911 reports) (2012)
Social Disorder and Crime (911 Reports)
Physical Disorder (311 reports) (2015)
Usage of 311 System (2015)
Gentrification (2000 - 2014)
American Community Survey (2010-2014)
☑ Basic Characteristics, ACS 2010-2014
0.0 - 0.14
0.14 - 0.17
0.17 - 0.21
0.21 - 0.25
0.25 - 0.31
0.31 - 0.98
☐ Racial and Ethnic Composition, ACS 20
☐ Economic Characteristics, ACS 2010-20
☐ Highest Level of Education, ACS 2010-2
☐ Family and Household Characteristics, A
☐ Transportation to Work, ACS 2010-2014
☐ Housing Characteristics, ACS 2010-201
American Community Survey (2011-2015)
☐ Basic Characteristics, ACS 2011-2015
☐ Racial and Ethnic Composition, ACS 20
Census 2010
Health & Human Ecology
Society & Demographics
☐ High School Drop Out Rate (by District)
Historical Census
Historic Maps
Winter Flounder Spawning Closure Area
Quincy Bay
1 : 136494

The Dataverse Project

Archeology Example



Harvard Dataverse

Harvard Dataverse > Replication Data for: The Hydraulic Landscape of Nimrud

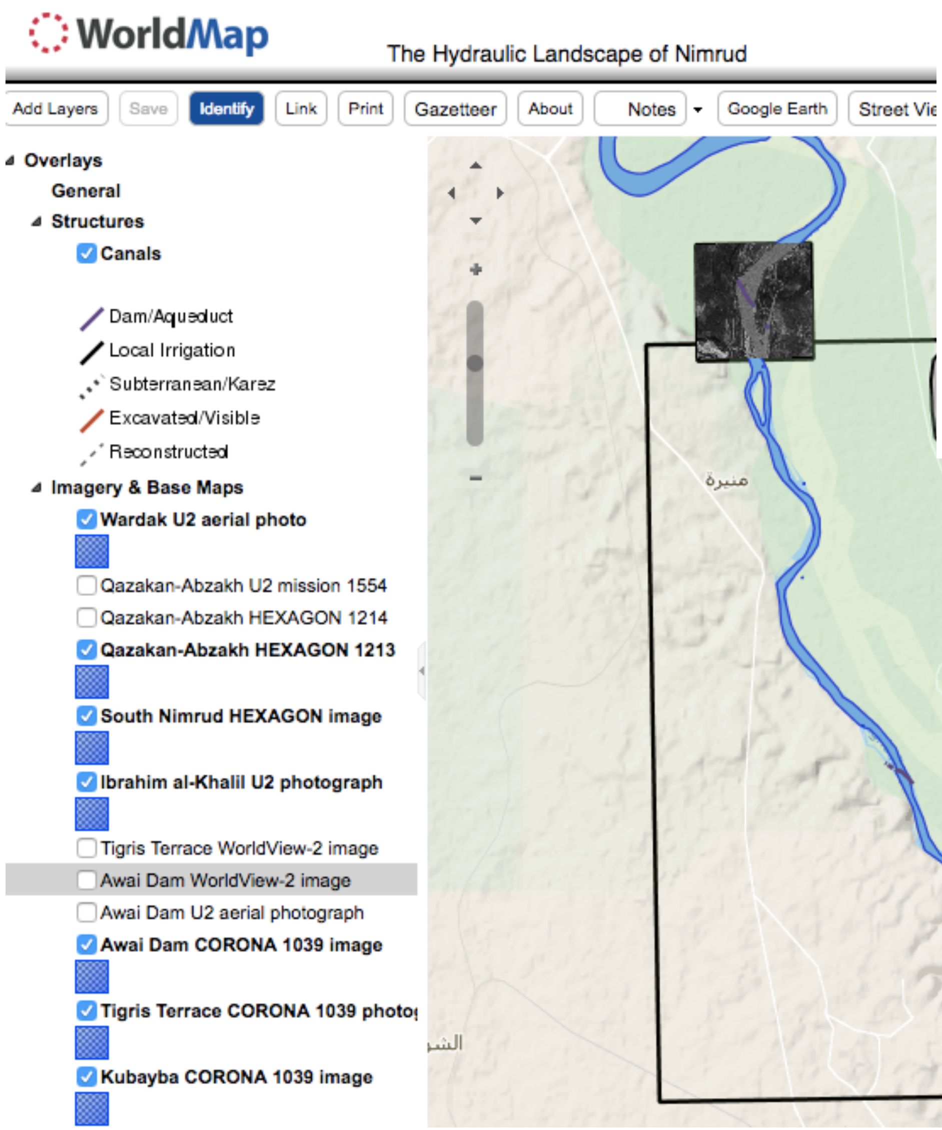
Replication Data for: The Hydraulic Landscape of Nimrud

Ur, Jason, 2015, "Replication Data for: The Hydraulic Landscape of Nimrud", doi:10.7910/DVN/W04NWB, Harvard Dataverse, V1

Description	Geospatial data for ancient sites and landscape features in the area of Nimrud, northern Iraq.
Subject	Social Sciences

2 Files

□	nimrud_canals.cpg Unknown - 5 bytes - Dec 16, 2015 - 3 Downloads MD5: ae3b3df9970b49b6523e608759bc957d	Download
□	nimrud_canals.zip Shapefile as ZIP Archive - 9.9 KB - Dec 16, 2015 - 3 Downloads MD5: 40ae0d5e97fcae97aa8a80b238d4abc9	Map Data Download



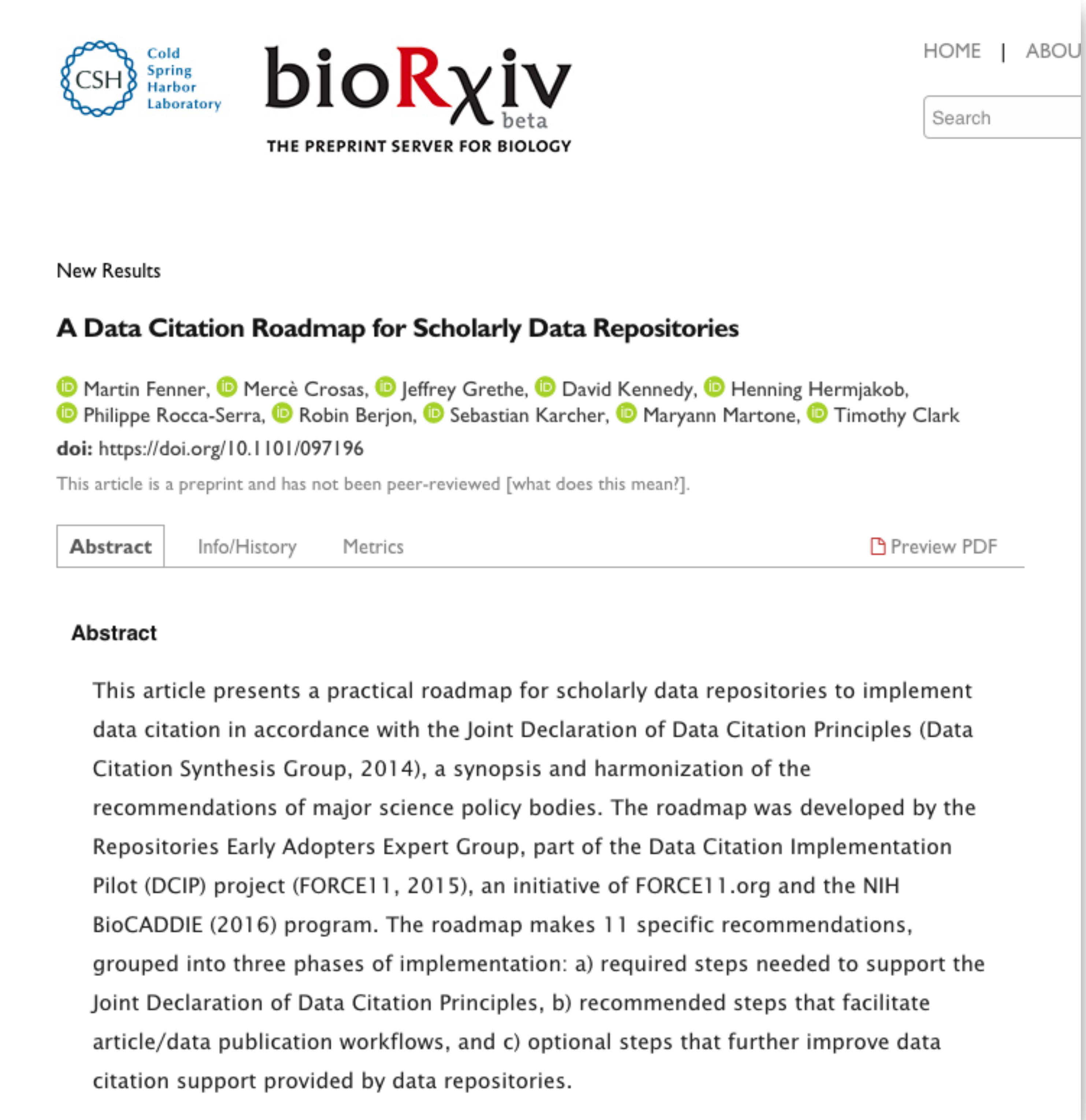
- Overlays
  - General
  - Structures
    - ☑ Canals
      - Dam/Aqueduct
      - Local Irrigation
      - Subterranean/Karez
      - Excavated/Visible
      - Reconstructed
  - Imagery & Base Maps
    - ☑ Wardak U2 aerial photo
    - □ Qazakan-Abzakh U2 mission 1554
    - □ Qazakan-Abzakh HEXAGON 1214
    - ☑ Qazakan-Abzakh HEXAGON 1213
    - ☑ South Nimrud HEXAGON image
    - ☑ Ibrahim al-Khalil U2 photograph
    - □ Tigris Terrace WorldView-2 image
    - □ Awai Dam WorldView-2 image
    - □ Awai Dam U2 aerial photograph
    - ☑ Awai Dam CORONA 1039 image
    - ☑ Tigris Terrace CORONA 1039 photo
    - ☑ Kubayba CORONA 1039 image

In the works: data citation roadmap to improve data discoverability

- Force11 Data Citation Implementation Pilot

- Landing page for dataset with machine-actionable standard citation metadata

- Working with Google to include dataset metadata in schema.org


CSH
Cold Spring Harbor Laboratory
bioRχiv beta
THE PREPRINT SERVER FOR BIOLOGY
HOME | ABOU
Search

New Results

A Data Citation Roadmap for Scholarly Data Repositories

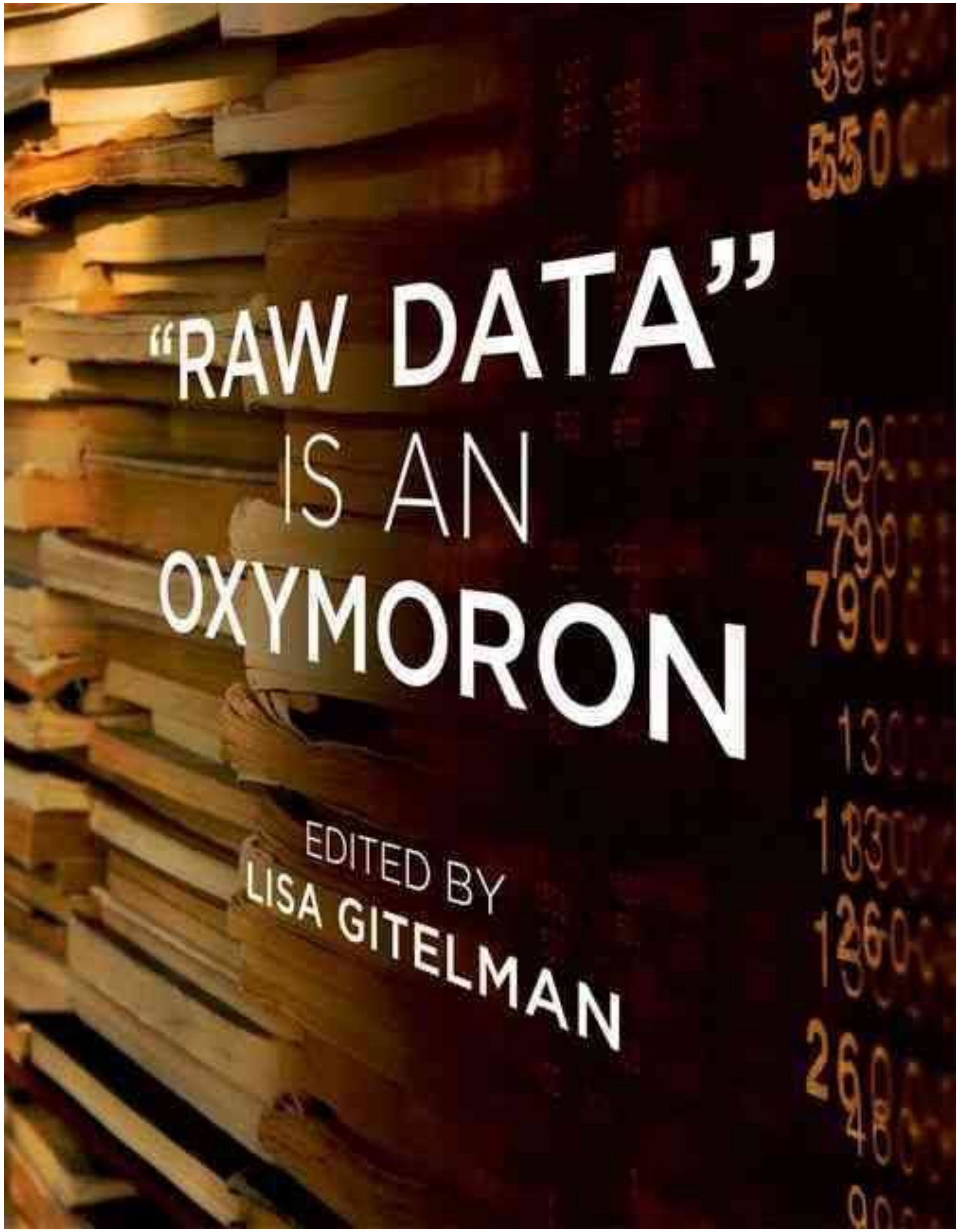
Martin Fenner, Mercè Crosas, Jeffrey Grethe, David Kennedy, Henning Hermjakob, Philippe Rocca-Serra, Robin Berjon, Sebastian Karcher, Maryann Martone, Timothy Clark

doi: https://doi.org/10.1101/097196

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract

This article presents a practical roadmap for scholarly data repositories to implement data citation in accordance with the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group, 2014), a synopsis and harmonization of the recommendations of major science policy bodies. The roadmap was developed by the Repositories Early Adopters Expert Group, part of the Data Citation Implementation Pilot (DCIP) project (FORCE11, 2015), an initiative of FORCE11.org and the NIH BioCADDIE (2016) program. The roadmap makes 11 specific recommendations, grouped into three phases of implementation: a) required steps needed to support the Joint Declaration of Data Citation Principles, b) recommended steps that facilitate article/data publication workflows, and c) optional steps that further improve data citation support provided by data repositories.



"we shouldn't think of data as a natural resource but as a cultural one that needs to be generated, protected, and interpreted."

Thank you!

T: @mercecrosas W: mercecrosas.com