Data Federation

Standards, Best Practices, and Thoughts

Sixth Open Research Cloud Alliance Workshop, May 22, 2019

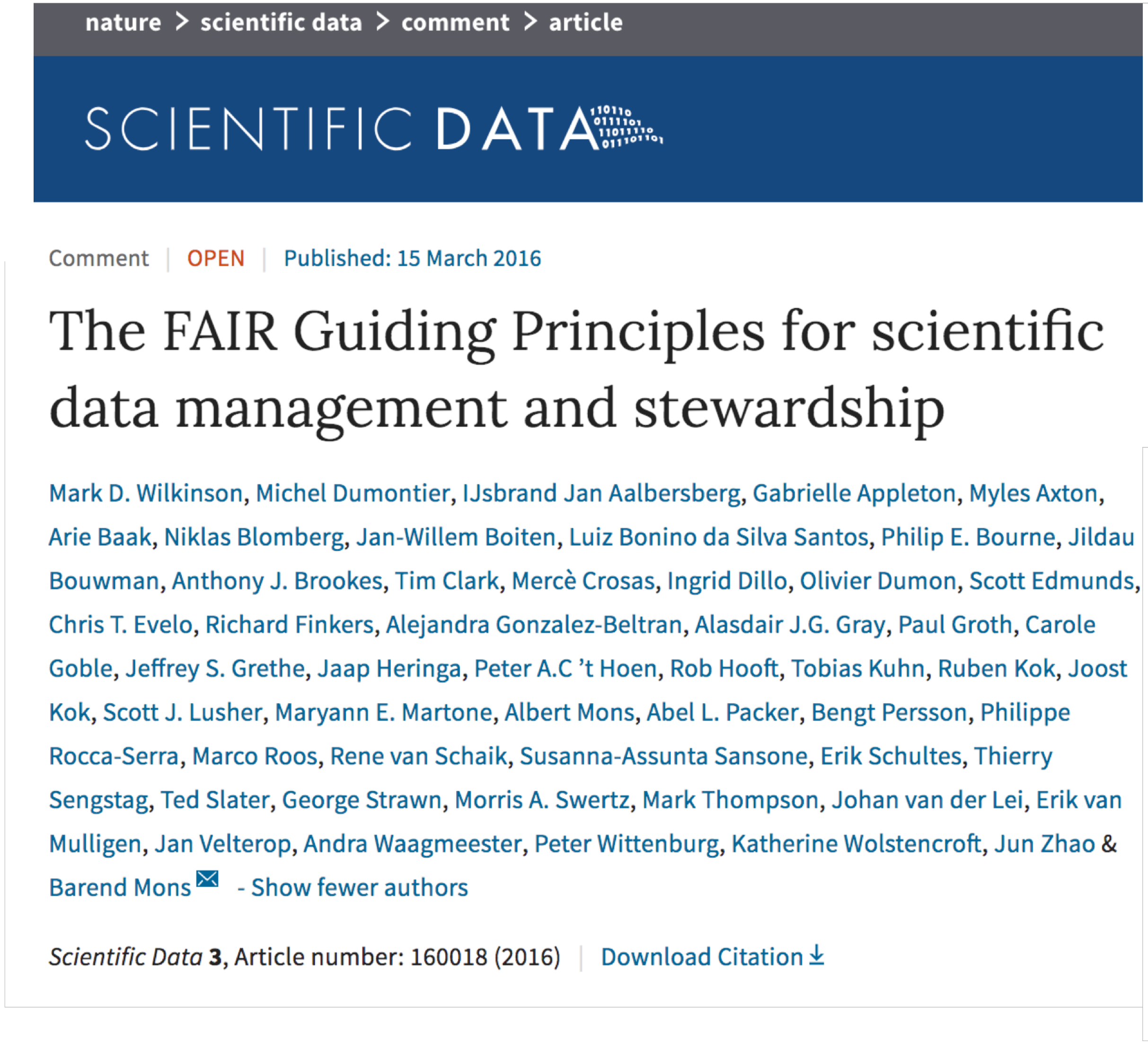
Mercè Crosas, Ph.D.

Harvard University's Research Data Officer, Office of Vice Provost for Research

Chief Data Science and Technology Officer, Institute for Quantitative Social Science

1. FAIR principles for data sharing
2. Two Scenarios for data federation
3. Data federation in Dataverse
4. Standards for data privacy and access requirements
5. Other considerations and uses cases

Three Years Ago ...



nature > scientific data > comment > article

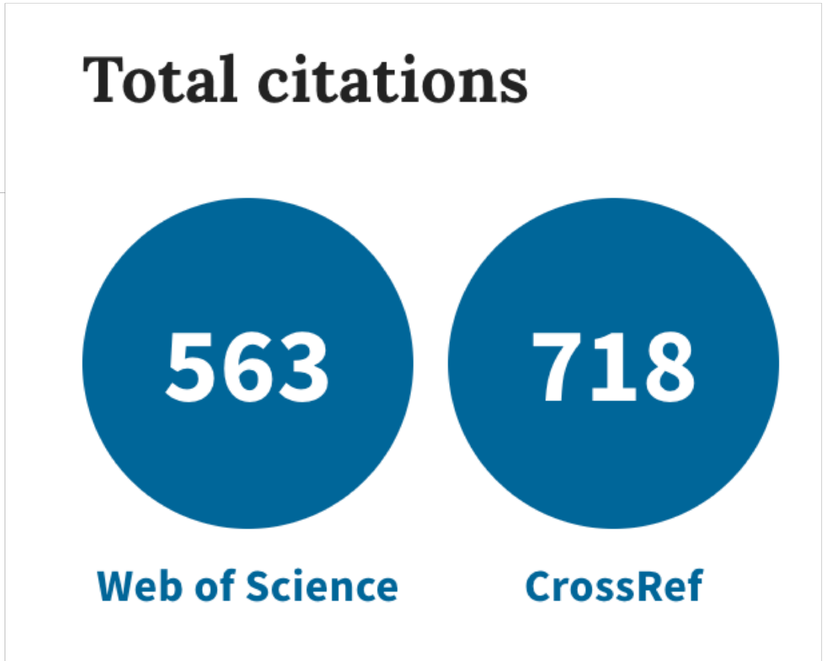
SCIENTIFIC DATA

Comment | OPEN | Published: 15 March 2016

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

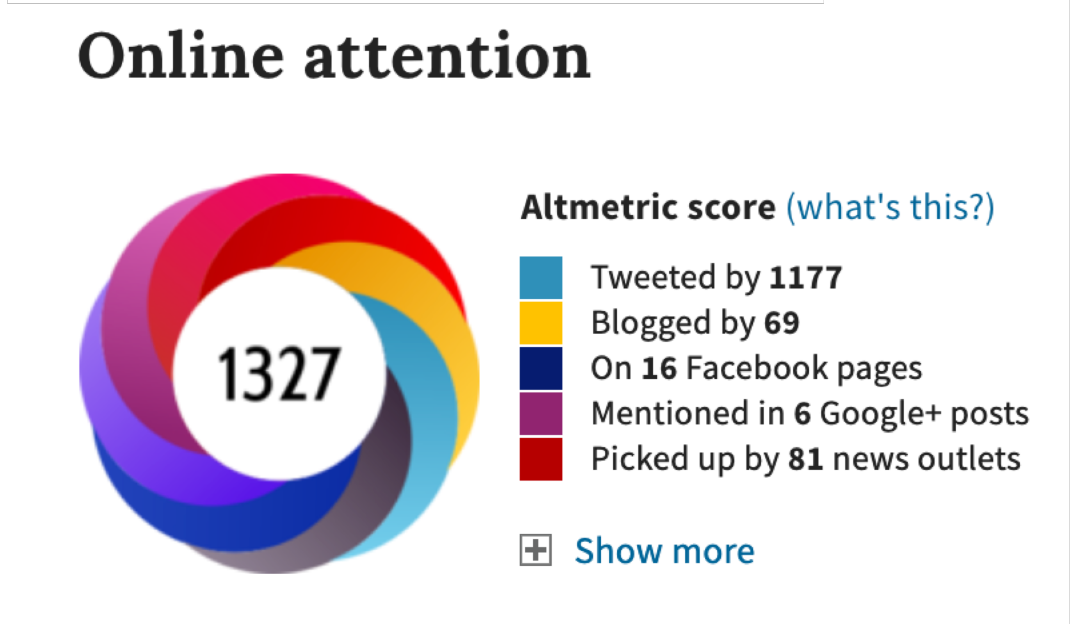
Scientific Data **3**, Article number: 160018 (2016) | Download Citation ⤓



Total citations

563 Web of Science

718 CrossRef



Online attention

Altmetric score (what's this?)

- 🟦 Tweeted by **1177**
- 🟨 Blogged by **69**
- 🟦 On **16** Facebook pages
- 🟪 Mentioned in **6** Google+ posts
- 🟥 Picked up by **81** news outlets

⊞ Show more

”The 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. ”

Wilkinson et al. 2016 “The FAIR Guiding Principles for scientific data management and stewardship” Scientific Data

The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

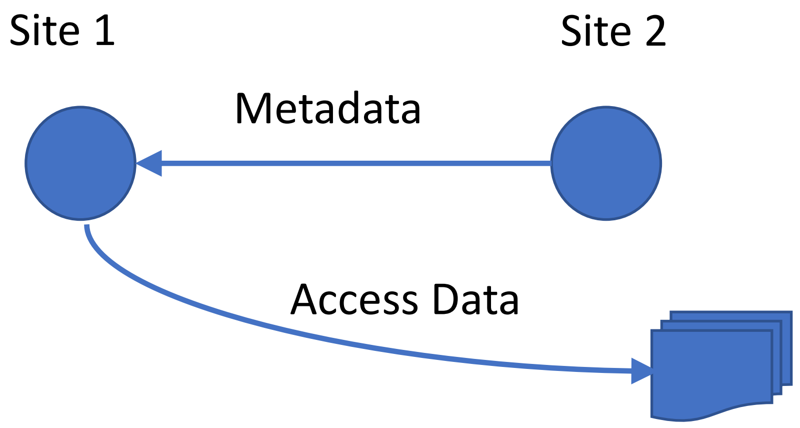
The FAIR Guiding Principles: Key Elements

- The importance of global persistent identifiers
- Metadata is essential
- Use of open, free, standard protocols
- Authentication & authorization and clear licenses, when needed

1. FAIR principles for data sharing
2. Two Scenarios for data federation
3. Data federation in Dataverse
4. Standards for data privacy and access requirements
5. Other considerations and use cases

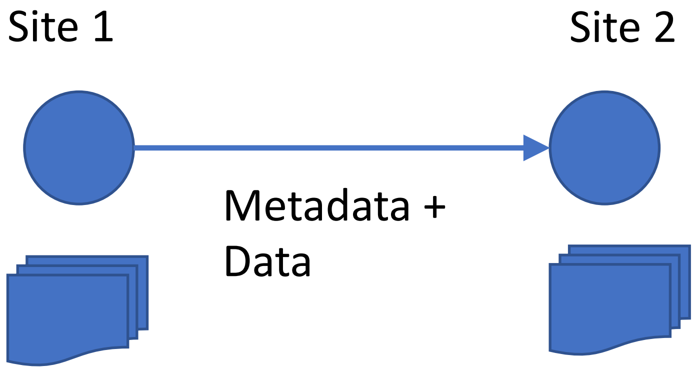
Scenario 1: Metadata Harvesting

- Metadata is harvested from one data site to another
- Data are accessed in the original remote site



Scenario 2: Metadata + Data are copied

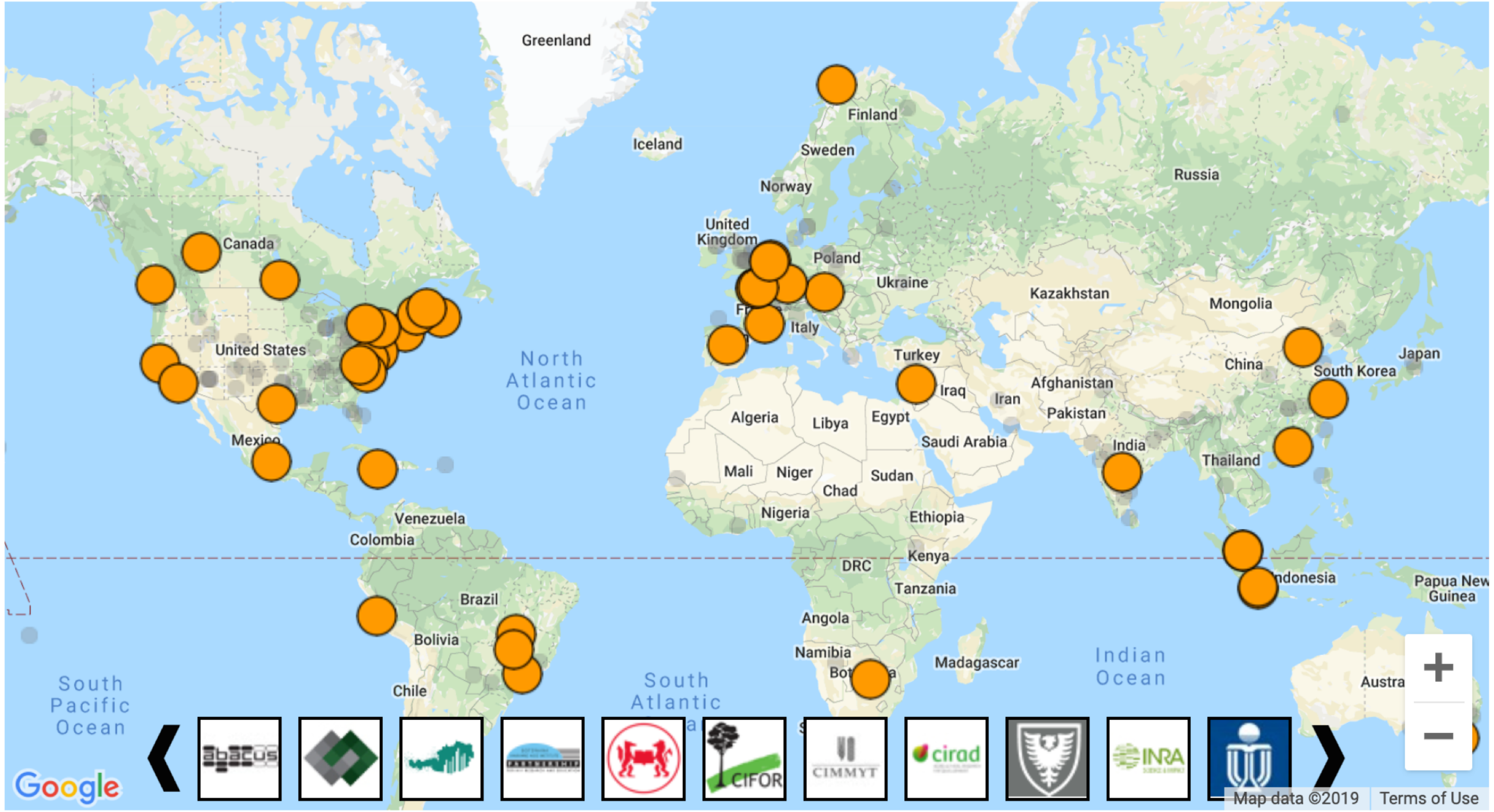
- Metadata and data are copied from one data site to another
- Data are accessed locally



1. FAIR principles for data sharing
2. Two Scenarios for data federation
3. Data federation in Dataverse
4. Standards for data privacy and access requirements
5. Other considerations and use cases

The Dataverse® Project

Open source research data repository software



A software, a community, many repositories

The Dataverse Software

http://dataverse.org

- Developed since 2006 at Harvard's Institute for Quantitative Social Science
- 89 contributors, most external to Harvard
- > 1000 pull requests in GitHub
- 12 releases a year
- 43 installations around the world

The Harvard Dataverse Repository

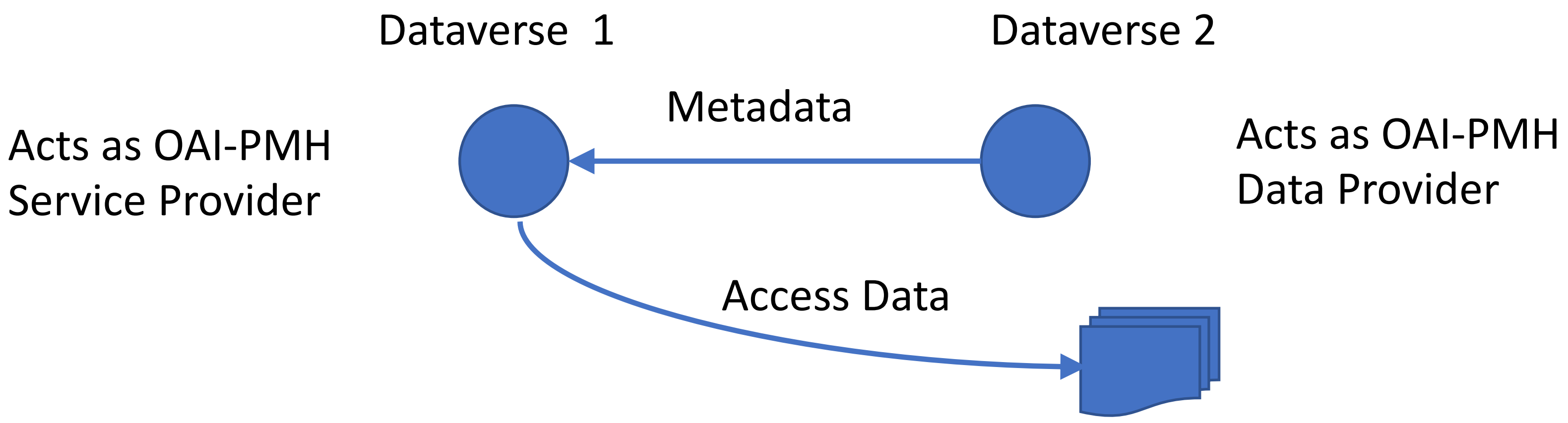
http://dataverse.harvard.edu

- Open to all researchers, all disciplines
- 30,000 datasets deposited
- + 50,000 datasets harvested from other Repositories
- 250 new datasets added per month
- 7 million file downloads

A Rich Set of Features, aligned with FAIR Principles

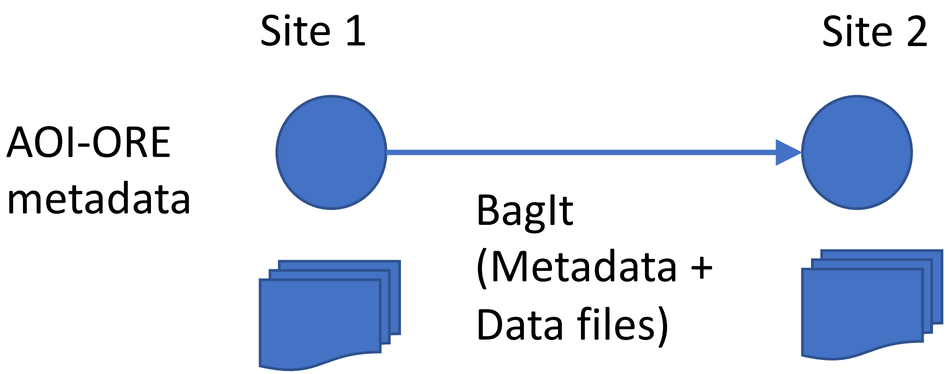
- Data citation with DOI: credit as an incentive to share data
- Metadata: find and reuse data
  - Data Documentation Initiative (DDI)
  - DataCite (+ OpenAire)
  - Dublin Core
  - Schema.org
- Versioning for dataset and files
- Tiered access to data: guestbook, terms of use and licenses, file restrictions
- APIs and Integration with data tools
- Customization and branding of your own dataverse (your collection of datasets)

Metadata Harvesting in Dataverse



- Uses Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH)
- Exposes metadata standards in XML format:
  - Dublin Core, Data Documentation Initiative (DDI)
- Widely used for interoperability with other repositories (e.g., OpenAire)

Copy Metadata + Data in Dataverse



Site 1

Site 2

AOI-ORE metadata

BagIt (Metadata + Data files)

- Uses Open Archive Initiative Object Reuse and Exchange (OAI-ORE):
    - Standards for the description and exchange of aggregations of resources
- Describes all files in a dataset with rich metadata in JSON-LD (RDF based)
- Uses BagIt to package all files (each referenced using DOIs) + metadata in a dataset

Recently architected and developed by Jim Myers, Senior Developer, Qualitative Data Repository (QDR). DataOne also uses it.

1. FAIR principles for data sharing
2. Two Scenarios for data federation
3. Data federation in Dataverse
4. Standards for data privacy and access requirements
5. Other considerations and use cases

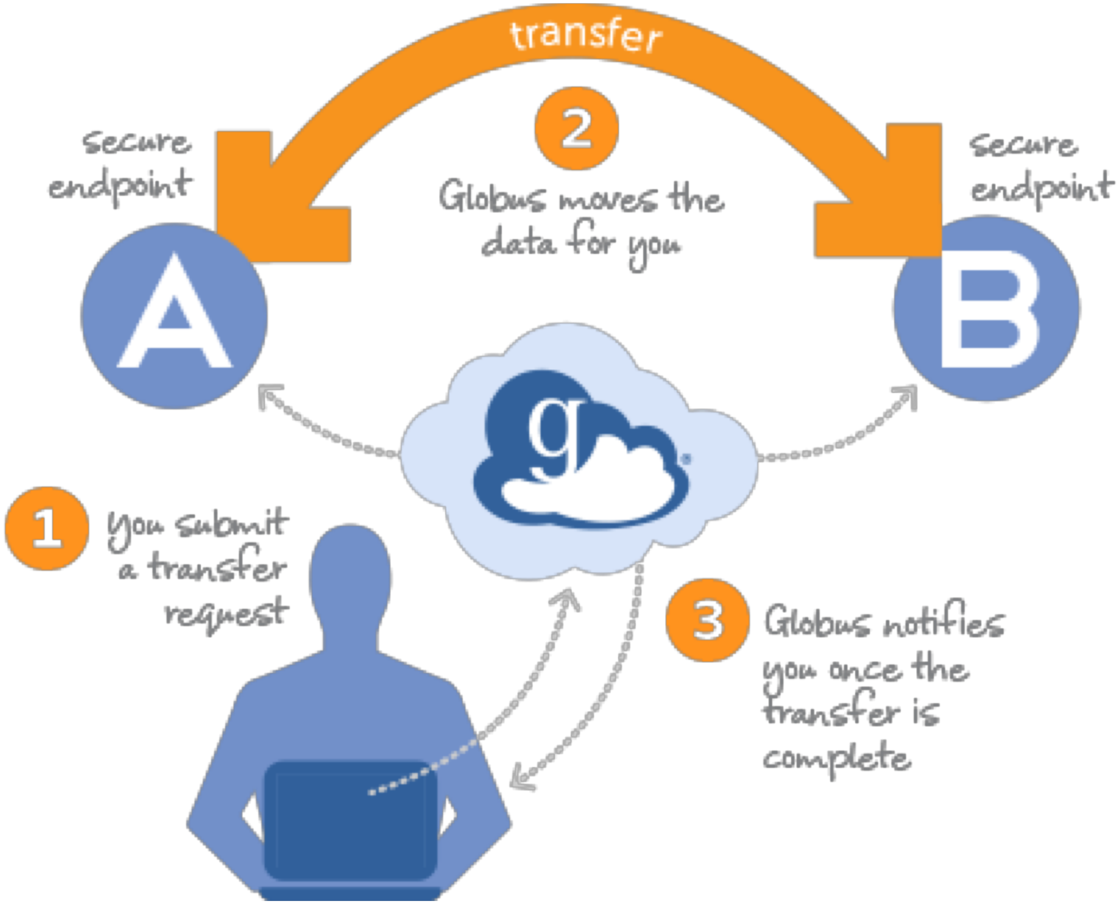
DataTags Facilitate Sharing Sensitive Data Responsibly

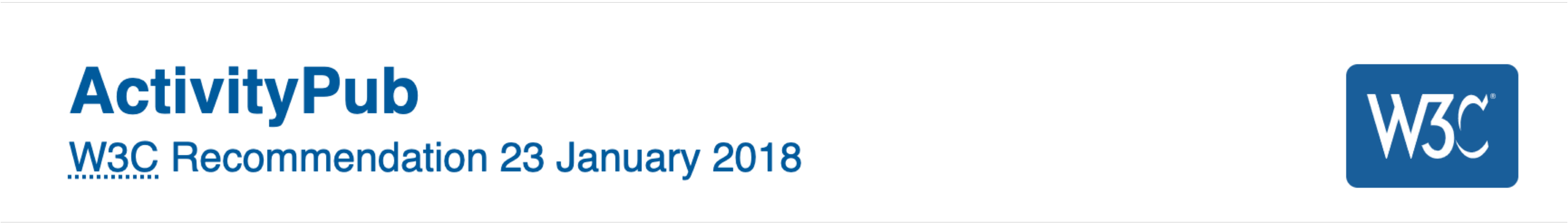
Blue	Public			
Green	Public Accountable	Register		
Yellow	Restricted Not Sensitive	Approval Needed	Click-thru Data Use Agreement (DUA)	Encrypted transmit
Orange	Restricted Sensitive	Approval Needed	Signed DUA	Encrypted transmit Encrypted storage
Red	Restricted High Sensitive	Approval Needed	Signed DUA Two-factor Auth	Encrypted transmit Encrypted storage
Crimson	Restricted Max Sensitive	Approval Needed	Signed DUA Two-factor Auth	Encrypted transmit Multi-encrypted storage

Sweeney, Crosas, Bar-Sinai, 2015. Sharing Sensitive Data with Confidence: The DataTags System, Technology Science

1. FAIR principles for data sharing
2. Two Scenarios for data federation
3. Data federation in Dataverse
4. Standards for data privacy and access requirements
5. Other considerations and use cases

Globus Transfer


transfer
2
Globus moves the data for you
secure endpoint
A
secure endpoint
B
g
1
you submit a transfer request
3
Globus notifies you once the transfer is complete

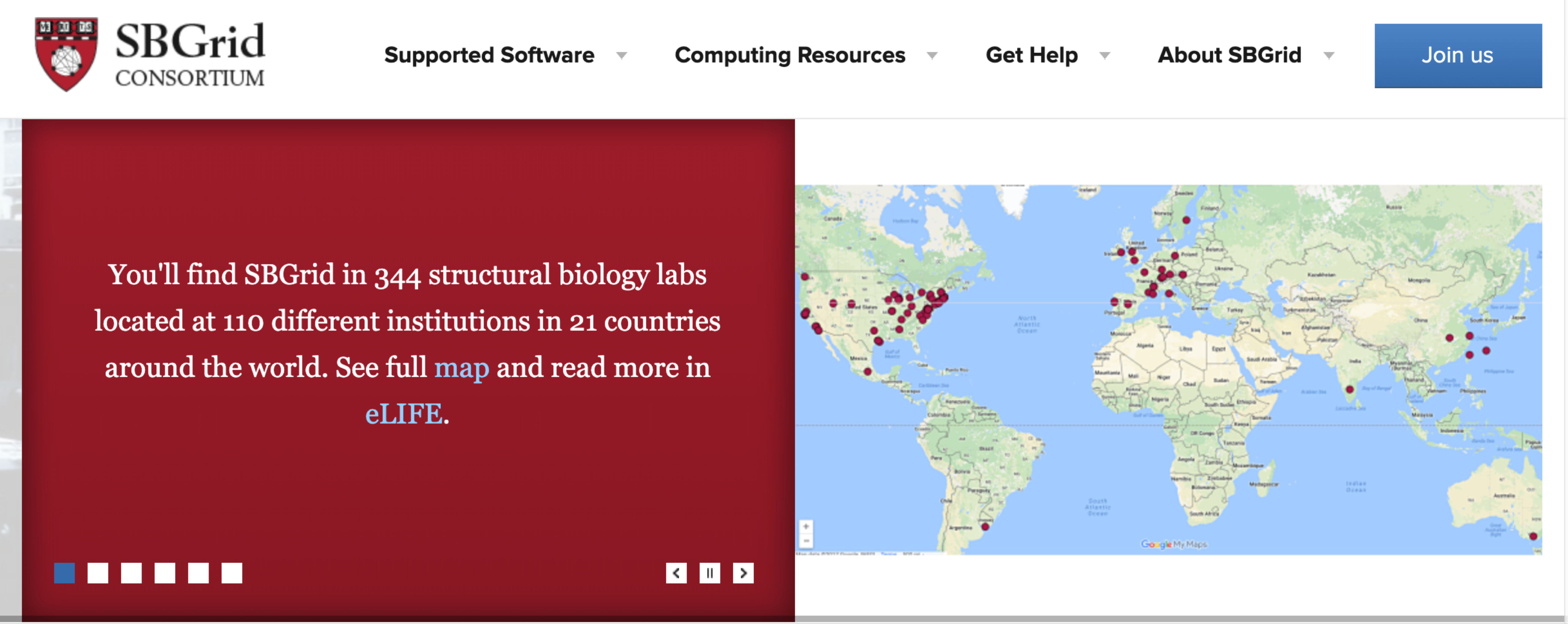


ActivityPub

W3C Recommendation 23 January 2018

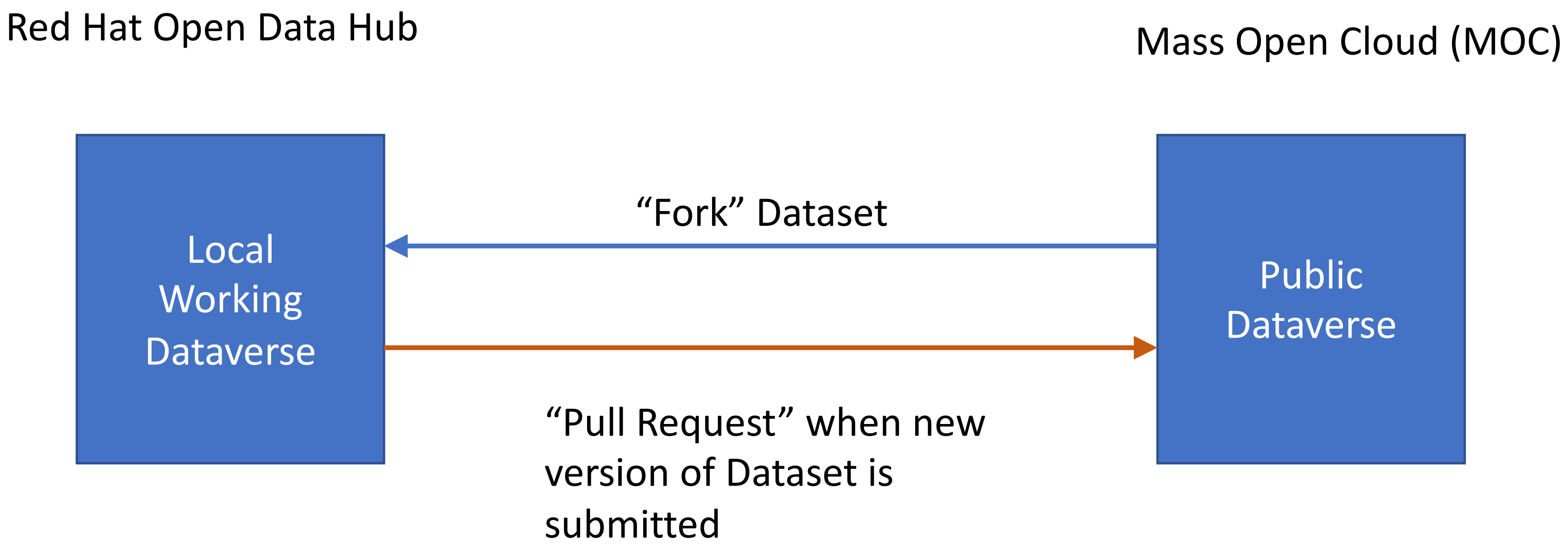
“The ActivityPub protocol is a decentralized social networking protocol based upon the [ActivityStreams] 2.0 data format. It provides a client to server API for creating, updating and deleting content, as well as a federated server to server API for delivering notifications and content.”

Use Case: SBGrid Data



You'll find SBGrid in 344 structural biology labs located at 110 different institutions in 21 countries around the world. See full map and read more in eLIFE.

Use Case: Dataverse, MOC, Open Data Hub



Red Hat Open Data Hub

Mass Open Cloud (MOC)

Local Working Dataverse

“Fork” Dataset

Public Dataverse

“Pull Request” when new version of Dataset is submitted

1. FAIR principles for data sharing
2. Two Scenarios for data federation
3. Data federation in Dataverse
4. Standards for data privacy and access requirements
5. Other considerations

Thanks