The Dataverse Project Is Also A Home For Life Sciences Data

Sonia Barbosa & Eleni Castro, Harvard University
bioCADDIE webinar: June 8, 2016
http://dataverse.org

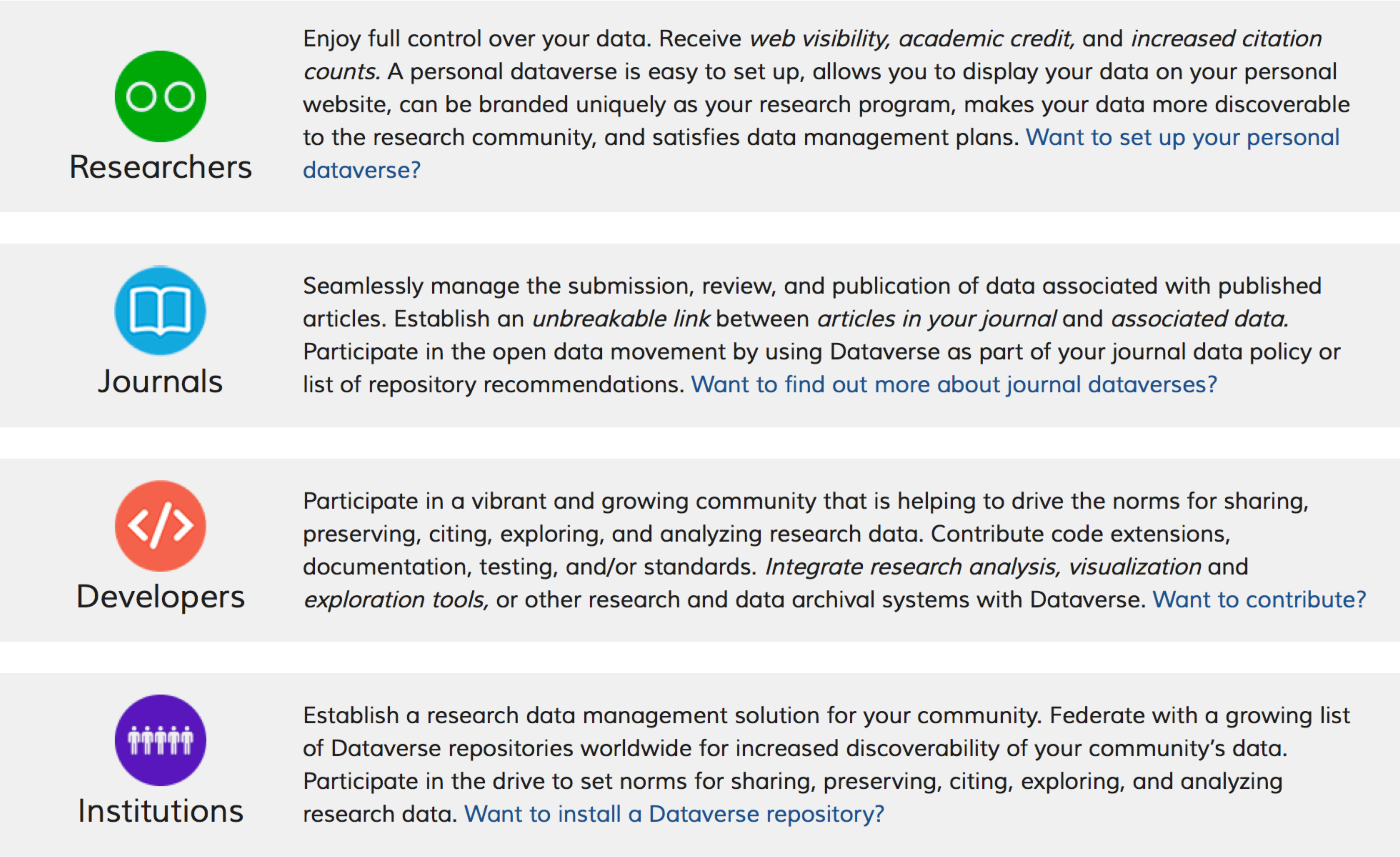
The Dataverse Project

Open source research data repository software

Share, preserve, cite, explore, & analyze data

Collaborations

- The Institute for Quantitative Social Science (IQSS)
- the Harvard University Library
- Harvard University Information Technology
- The Open Data Assistance Program at Harvard (a collaboration with Harvard Library, the Office for Scholarly Communication and IQSS)
- The Library Technology Services at HUIT provides hosting and backup support



Researchers

Enjoy full control over your data. Receive *web visibility, academic credit,* and *increased citation counts.* A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals

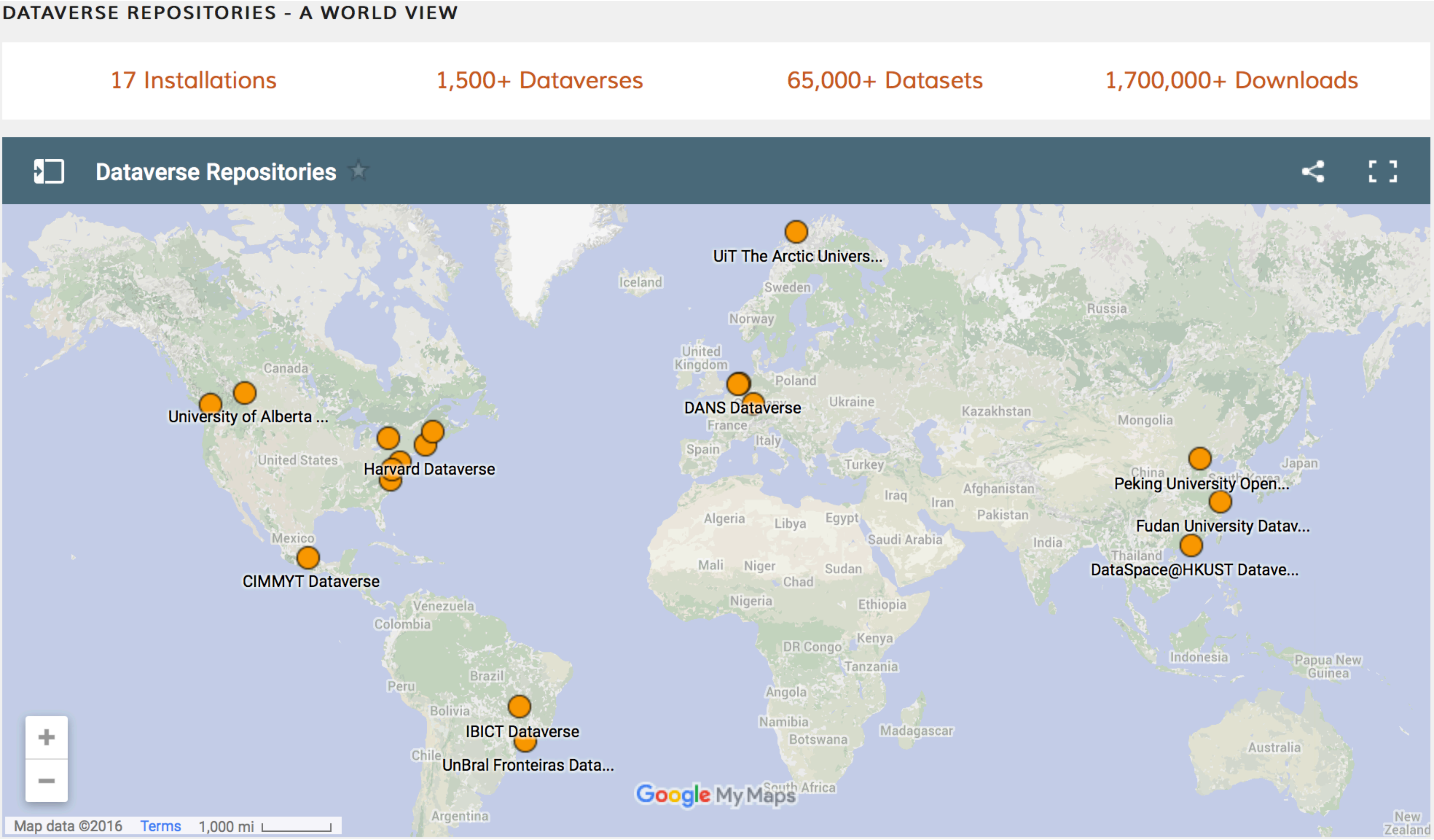
Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data.* Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Developers

Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *exploration tools,* or other research and data archival systems with Dataverse. Want to contribute?

Institutions

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?



DATAVERSE REPOSITORIES - A WORLD VIEW

17 Installations

1,500+ Dataverses

65,000+ Datasets

1,700,000+ Downloads

Features

Data Citation
automatically generated

Multiple Publishing Workflows
dataset in draft, in review, and then published

Terms of Use + Guestbook
CC0 waiver default, custom terms of use, and download metrics

Account + Data Notifications
access request, roles granted, and when data is published to name a few

Faceted Search
metadata fields based facets

Pull header metadata from Astronomy (FITS) files

APIs for interoperability
search API, data deposit API

Three Levels of Metadata
description/citation, domain-specific or custom fields, file metadata

Access Control Support
pre-defined and custom roles

Restricted Files + Ability to request access to restricted files
allow anyone, certain people, or no one to be able to download files

Customization of dataverses
branding, metadata based facets, sub-dataverses, featured dataverses

Re-format, Summary Statistics, and Analysis for Tabular Files
integration with TwoRavens

Mapping of Geospatial files
integration with WorldMap

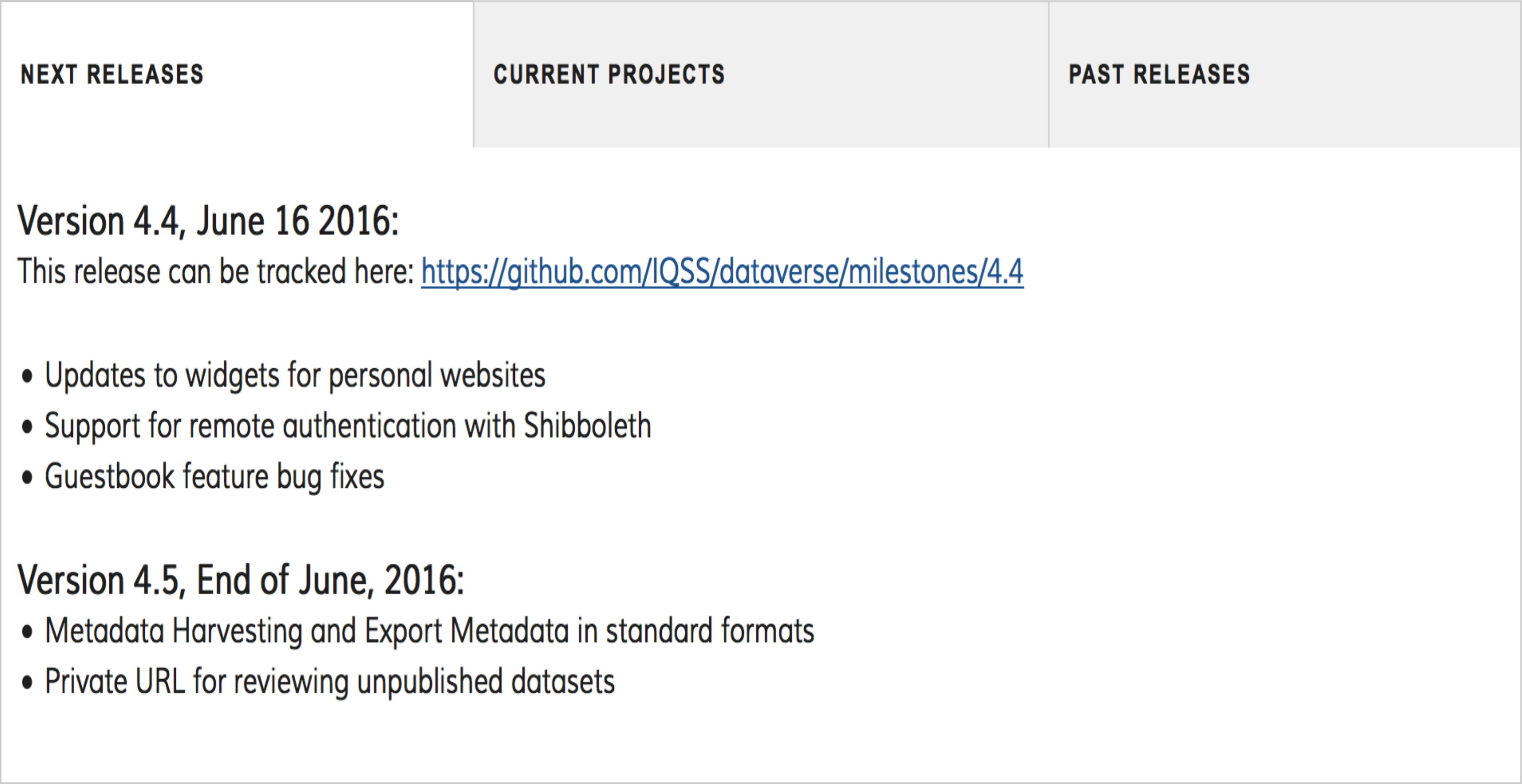
Features

- Standard Citation: Title, DOI, UNF, Versioning, Repository (following FORCE11 Joint Declaration of Data Citation Principles)
- File level support: MD5, UNF, Tabular data, multiple download options, tags, descriptions, zip extraction, audio, video, PDF preview, image files w/preview, unlimited files, all file types
- Metadata support
- Terms: CC0, additional terms, restricted/open, application forms
- Versioning

Features…

- dataverse or dataset
- Themes and widgets
- Permissions
- Groups
- Guestbook
- Templates
- Featured dataverses

Next Releases



Version 4.4, June 16 2016:

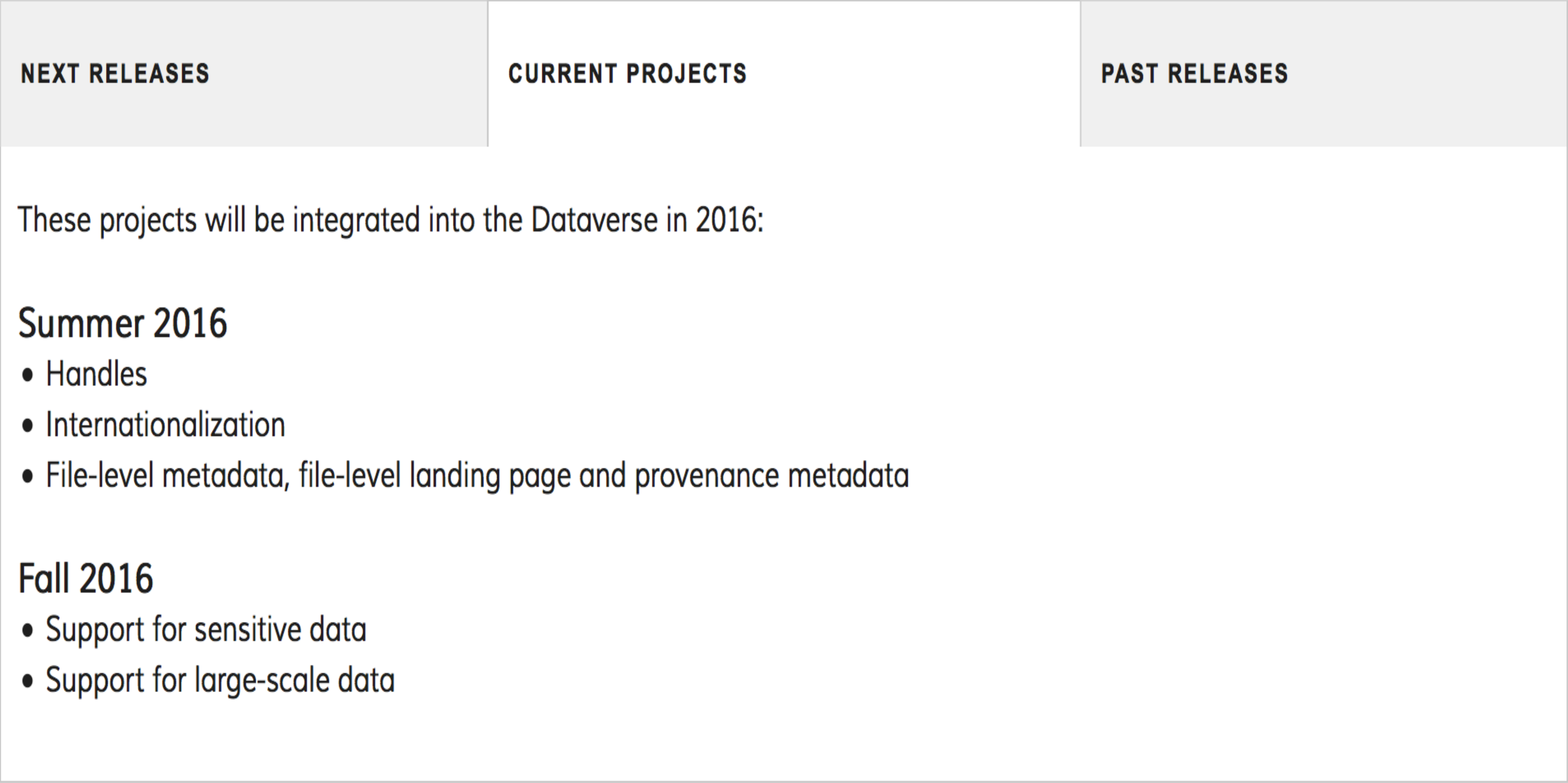
This release can be tracked here: https://github.com/IQSS/dataverse/milestones/4.4

- Updates to widgets for personal websites
- Support for remote authentication with Shibboleth
- Guestbook feature bug fixes

Version 4.5, End of June, 2016:

- Metadata Harvesting and Export Metadata in standard formats
- Private URL for reviewing unpublished datasets

Current Projects



These projects will be integrated into the Dataverse in 2016:

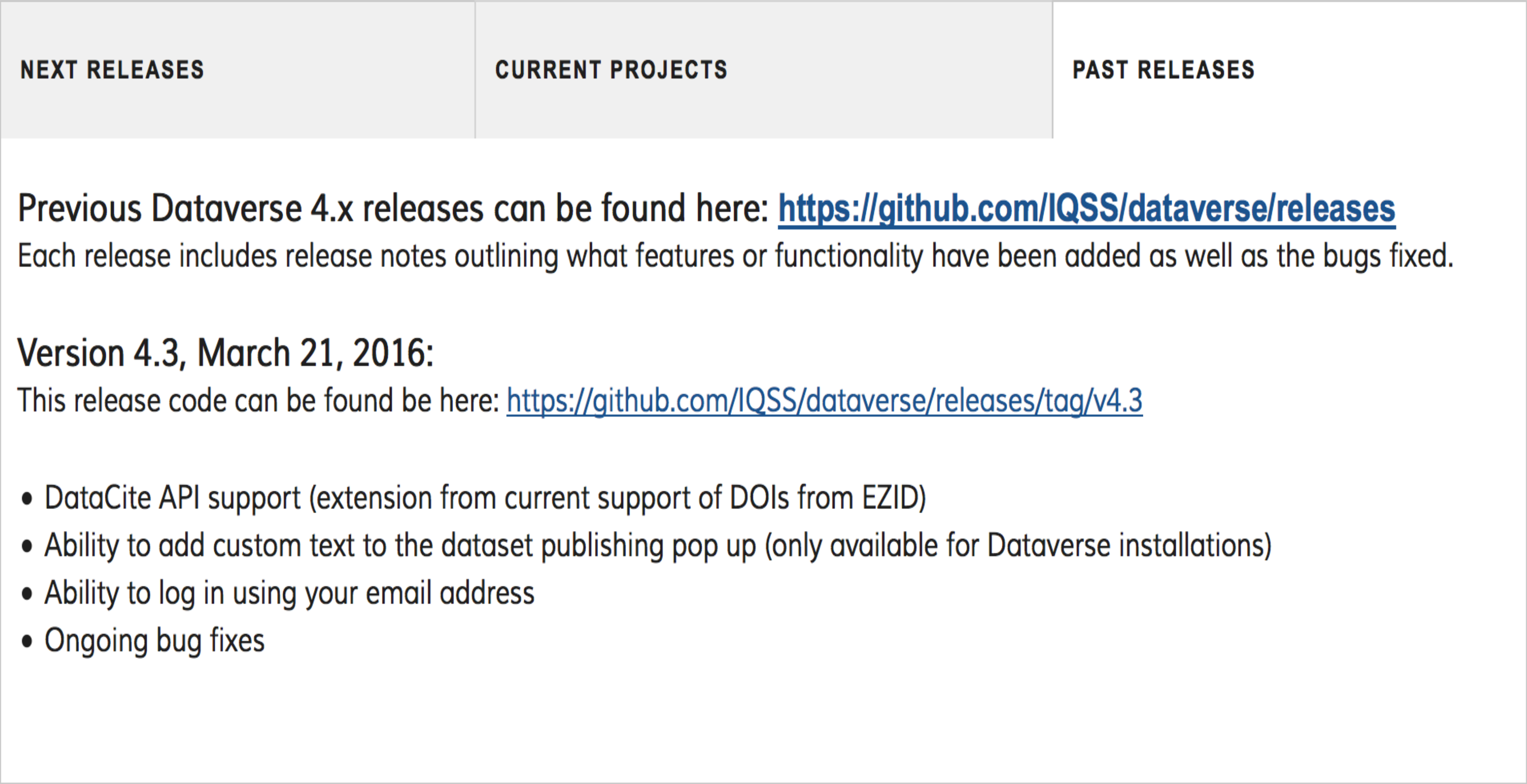
Summer 2016

- Handles
- Internationalization
- File-level metadata, file-level landing page and provenance metadata

Fall 2016

- Support for sensitive data
- Support for large-scale data

Past Releases



Previous Dataverse 4.x releases can be found here: https://github.com/IQSS/dataverse/releases

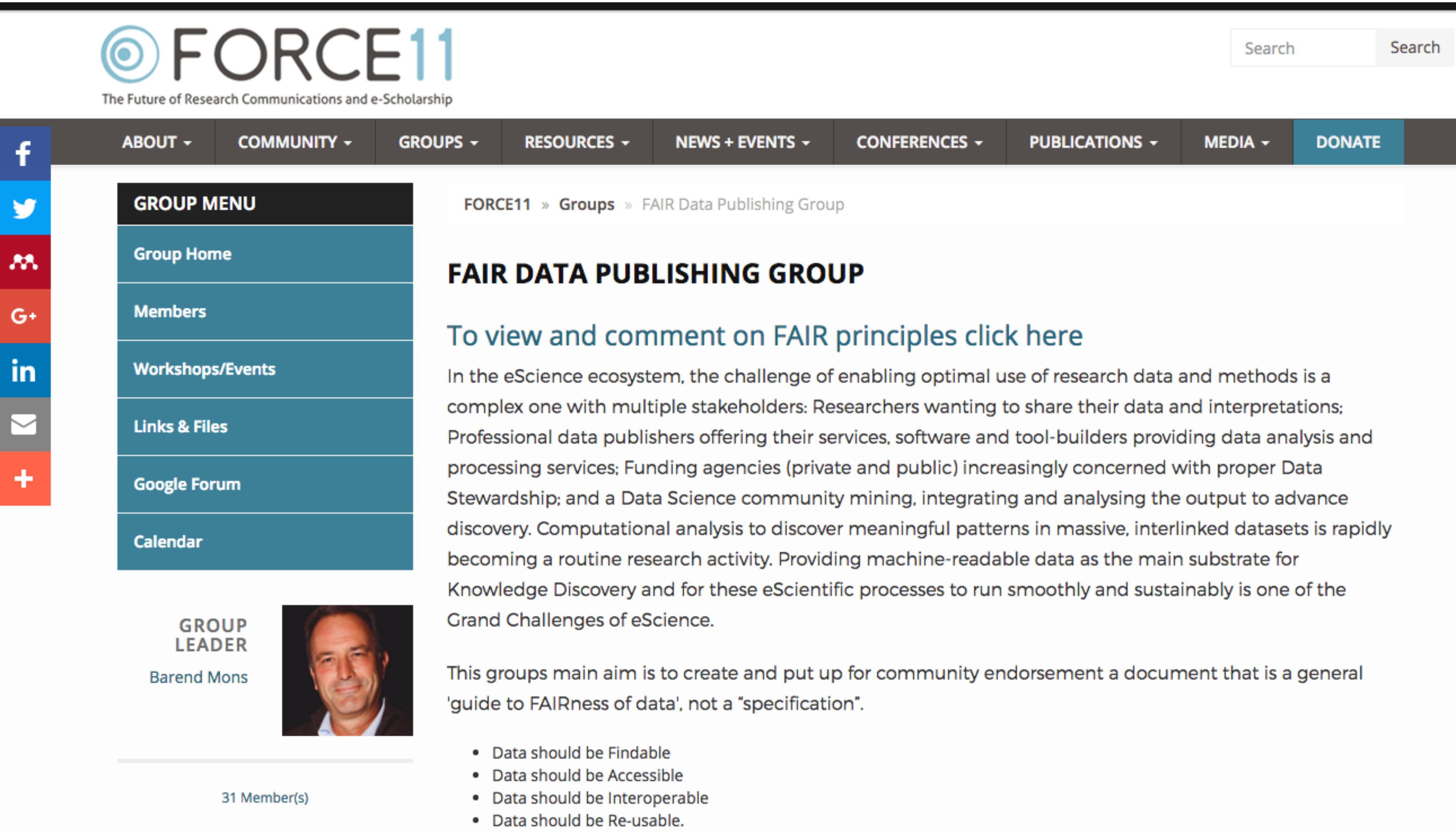
Each release includes release notes outlining what features or functionality have been added as well as the bugs fixed.

Version 4.3, March 21, 2016:

This release code can be found be here: https://github.com/IQSS/dataverse/releases/tag/v4.3

- DataCite API support (extension from current support of DOIs from EZID)
- Ability to add custom text to the dataset publishing pop up (only available for Dataverse installations)
- Ability to log in using your email address
- Ongoing bug fixes

Dataverse is working on being FAIR



FAIR DATA PUBLISHING GROUP

To view and comment on FAIR principles click here

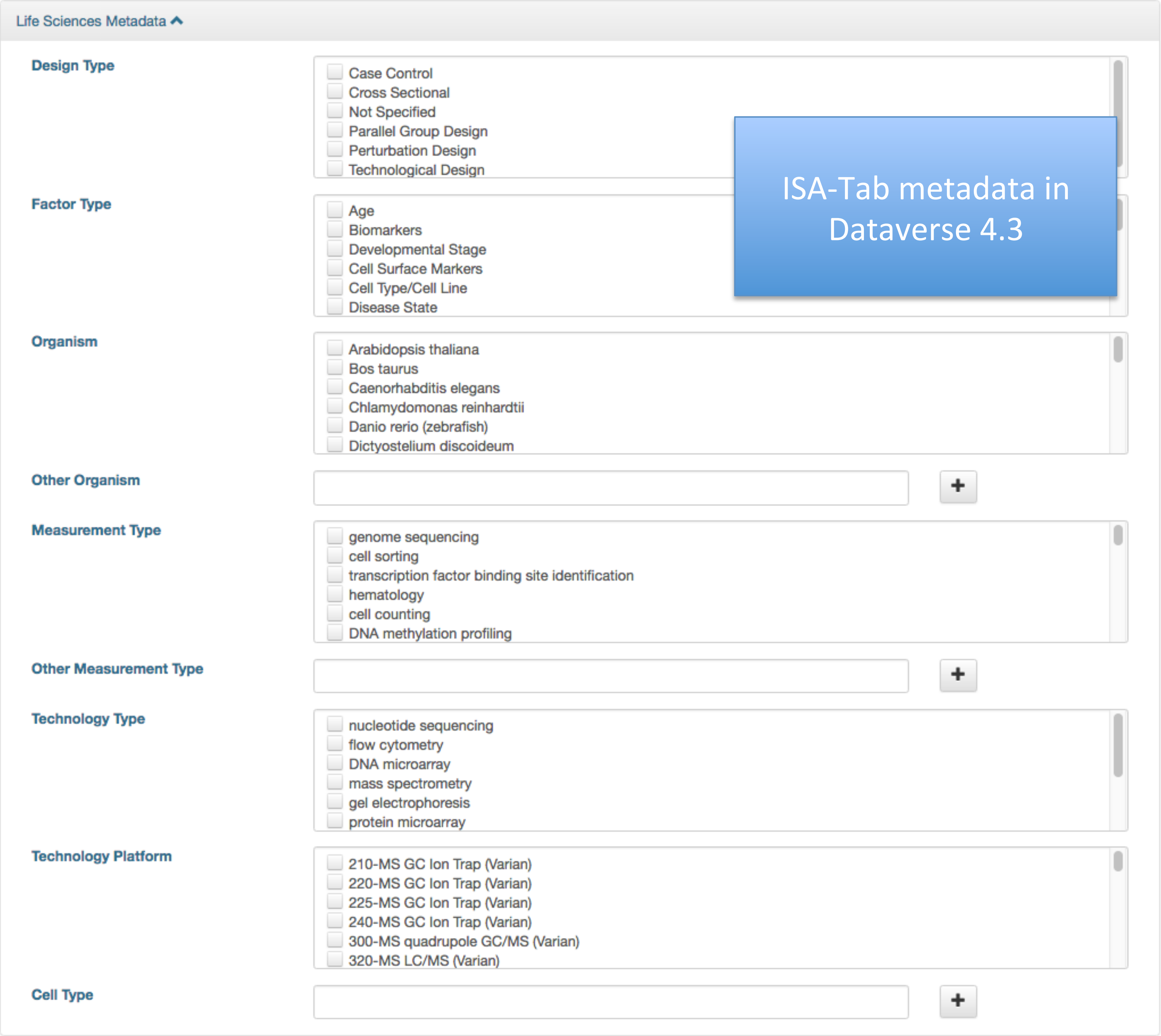
In the eScience ecosystem, the challenge of enabling optimal use of research data and methods is a complex one with multiple stakeholders: Researchers wanting to share their data and interpretations; Professional data publishers offering their services, software and tool-builders providing data analysis and processing services; Funding agencies (private and public) increasingly concerned with proper Data Stewardship; and a Data Science community mining, integrating and analysing the output to advance discovery. Computational analysis to discover meaningful patterns in massive, interlinked datasets is rapidly becoming a routine research activity. Providing machine-readable data as the main substrate for Knowledge Discovery and for these eScientific processes to run smoothly and sustainably is one of the Grand Challenges of eScience.

This groups main aim is to create and put up for community endorsement a document that is a general 'guide to FAIRness of data', not a “specification”.

- Data should be Findable
- Data should be Accessible
- Data should be Interoperable
- Data should be Re-usable.

Life Sciences Metadata

- ISA-Tab with Scientific Data flavor (see next slide)
- Various ontologies from bioportal (ex. OBI)
- NCBI Taxonomy
- Plan to support export (Fall 2016) – discoverability with NIH's discovery index



Life Sciences Metadata ^

Design Type

- □ Case Control
- □ Cross Sectional
- □ Not Specified
- □ Parallel Group Design
- □ Perturbation Design
- □ Technological Design

Factor Type

- □ Age
- □ Biomarkers
- □ Developmental Stage
- □ Cell Surface Markers
- □ Cell Type/Cell Line
- □ Disease State

Organism

- □ Arabidopsis thaliana
- □ Bos taurus
- □ Caenorhabditis elegans
- □ Chlamydomonas reinhardtii
- □ Danio rerio (zebrafish)
- □ Dictyostelium discoideum

Other Organism □ +

Measurement Type

- □ genome sequencing
- □ cell sorting
- □ transcription factor binding site identification
- □ hematology
- □ cell counting
- □ DNA methylation profiling

Other Measurement Type □ +

Technology Type

- □ nucleotide sequencing
- □ flow cytometry
- □ DNA microarray
- □ mass spectrometry
- □ gel electrophoresis
- □ protein microarray

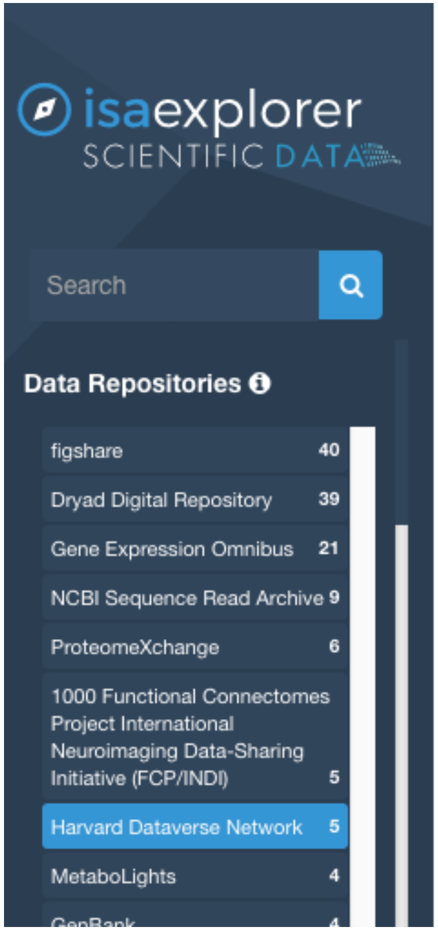
Technology Platform

- □ 210-MS GC Ion Trap (Varian)
- □ 220-MS GC Ion Trap (Varian)
- □ 225-MS GC Ion Trap (Varian)
- □ 240-MS GC Ion Trap (Varian)
- □ 300-MS quadrupole GC/MS (Varian)
- □ 320-MS LC/MS (Varian)

Cell Type □ +

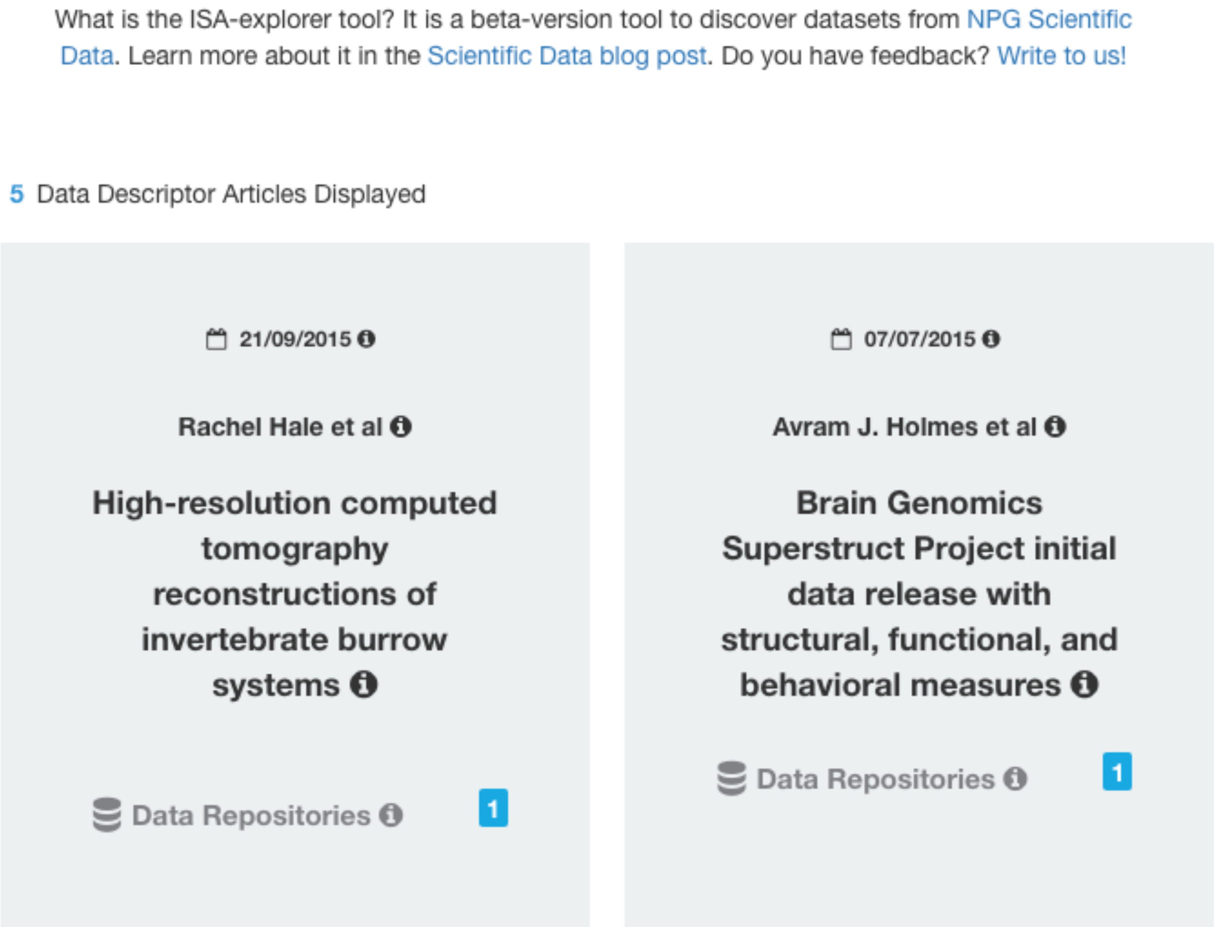
ISA-Tab metadata in Dataverse 4.3

Dataverse in isaexplorer



Data Repositories

figshare	40
Dryad Digital Repository	39
Gene Expression Omnibus	21
NCBI Sequence Read Archive	9
ProteomeXchange	6
1000 Functional Connectomes Project International Neuroimaging Data-Sharing Initiative (FCP/INDI)	5
Harvard Dataverse Network	5
MetaboLights	4
GenBank	4



What is the ISA-explorer tool? It is a beta-version tool to discover datasets from NPG Scientific Data. Learn more about it in the Scientific Data blog post. Do you have feedback? Write to us!

5 Data Descriptor Articles Displayed

21/09/2015

Rachel Hale et al

High-resolution computed tomography reconstructions of invertebrate burrow systems

Data Repositories 1

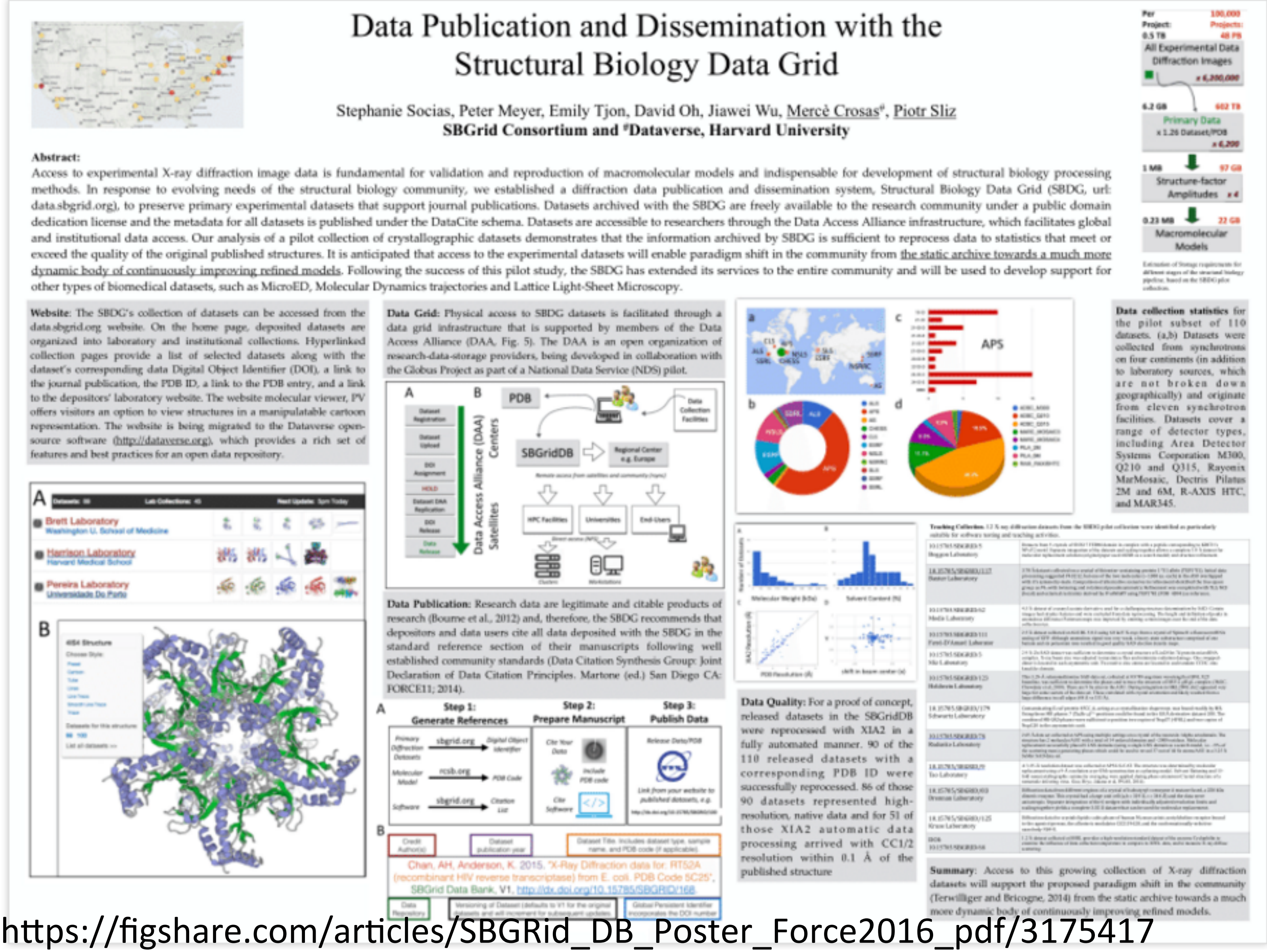
07/07/2015

Avram J. Holmes et al

Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures

Data Repositories 1

Structural Biology Data + Dataverse



Data Publication and Dissemination with the Structural Biology Data Grid

Stephanie Socias, Peter Meyer, Emily Tjon, David Oh, Jiawei Wu, Mercè Crosas[*], Piotr Sliz
SBGrid Consortium and [*]Dataverse, Harvard University

Abstract:
Access to experimental X-ray diffraction image data is fundamental for validation and reproduction of macromolecular models and indispensable for development of structural biology processing methods. In response to evolving needs of the structural biology community, we established a diffraction data publication and dissemination system, Structural Biology Data Grid (SBDG, url: data.sbgrid.org), to preserve primary experimental datasets that support journal publications. Datasets archived with the SBDG are freely available to the research community under a public domain dedication license and the metadata for all datasets is published under the DataCite schema. Datasets are accessible to researchers through the Data Access Alliance infrastructure, which facilitates global and institutional data access. Our analysis of a pilot collection of crystallographic datasets demonstrates that the information archived by SBDG is sufficient to reprocess data to statistics that meet or exceed the quality of the original published structures. It is anticipated that access to the experimental datasets will enable paradigm shift in the community from the static archive towards a much more dynamic body of continuously improving refined models. Following the success of this pilot study, the SBDG has extended its services to the entire community and will be used to develop support for other types of biomedical datasets, such as MicroED, Molecular Dynamics trajectories and Lattice Light-Sheet Microscopy.

Website: The SBDG's collection of datasets can be accessed from the data.sbgrid.org website. On the home page, deposited datasets are organized into laboratory and institutional collections. Hyperlinked collection pages provide a list of selected datasets along with the dataset's corresponding data Digital Object Identifier (DOI), a link to the journal publication, the PDB ID, a link to the PDB entry, and a link to the depositors' laboratory website. The website molecular viewer, PV offers visitors an option to view structures in a manipulatable cartoon representation. The website is being migrated to the Dataverse open-source software (http://dataverse.org), which provides a rich set of features and best practices for an open data repository.

Data Grid: Physical access to SBDG datasets is facilitated through a data grid infrastructure that is supported by members of the Data Access Alliance (DAA, Fig. 5). The DAA is an open organization of research-data-storage providers, being developed in collaboration with the Globus Project as part of a National Data Service (NDS) pilot.

Data Publication: Research data are legitimate and citable products of research (Bourne et al., 2012) and, therefore, the SBDG recommends that depositors and data users cite all data deposited with the SBDG in the standard reference section of their manuscripts following well established community standards (Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone (ed.) San Diego CA: FORCE11; 2014).

Data collection statistics for the pilot subset of 110 datasets. (a,b) Datasets were collected from synchrotrons on four continents (in addition to laboratory sources, which are not broken down geographically) and originate from eleven synchrotron facilities. Datasets cover a range of detector types, including Area Detector Systems Corporation M300, Q210 and Q315, Rayonix MarMosaic, Dectris Pilatus 2M and 6M, R-AXIS HTC, and MAR345.

Data Quality: For a proof of concept, released datasets in the SBGridDB were reprocessed with XIA2 in a fully automated manner. 90 of the 110 released datasets with a corresponding PDB ID were successfully reprocessed. 86 of those 90 datasets represented high-resolution, native data and for 51 of those XIA2 automatic data processing arrived with CC1/2 resolution within 0.1 Å of the published structure

Summary: Access to this growing collection of X-ray diffraction datasets will support the proposed paradigm shift in the community (Terwilliger and Bricogne, 2014) from the static archive towards a much more dynamic body of continuously improving refined models.

https://figshare.com/articles/SBGRid_DB_Poster_Force2016_pdf/3175417

Journals + Dataverse

SCIENTIFIC DATA

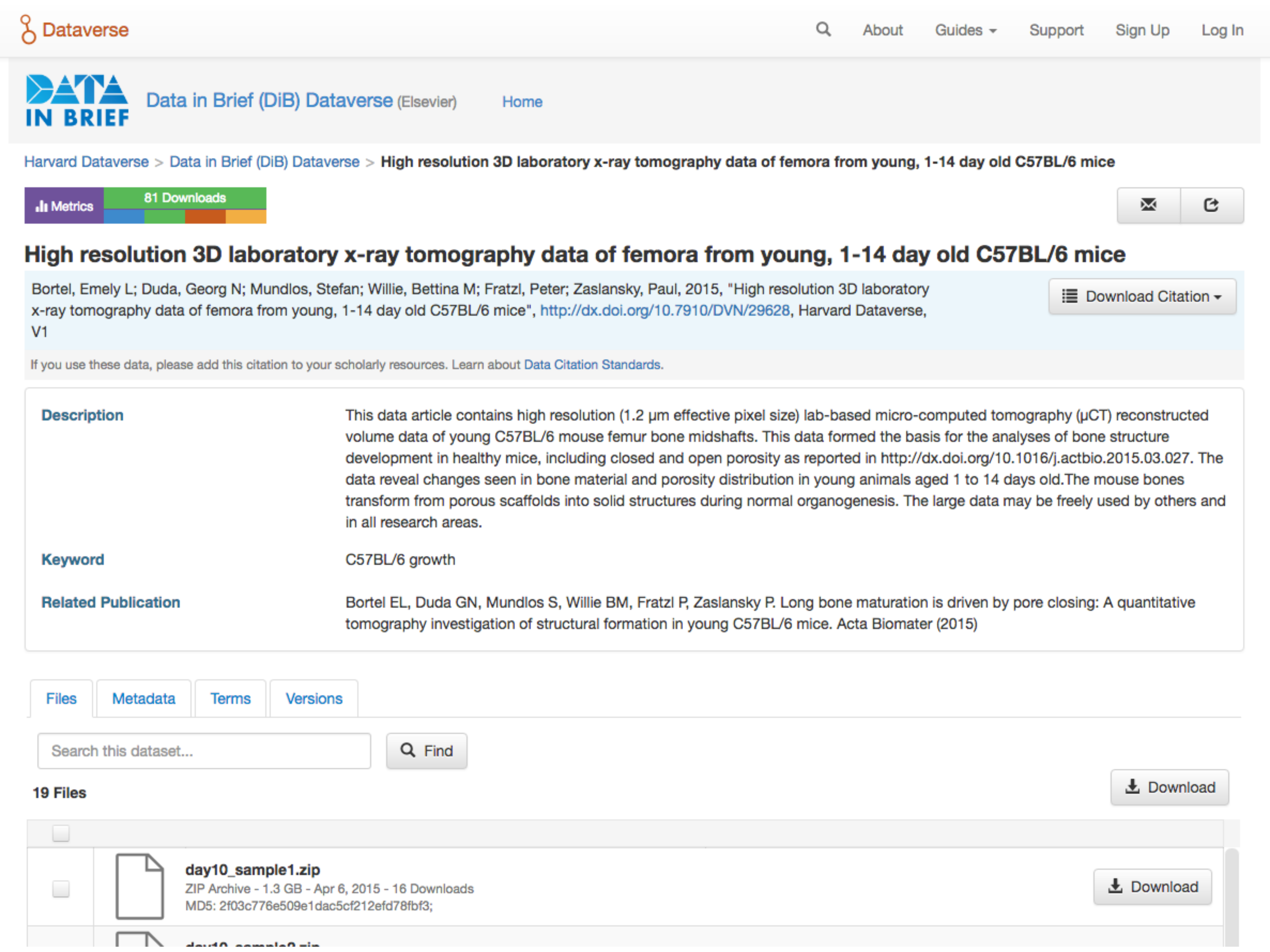
American Heart Association®

NON SOLUS

ELSEVIER

PLOS

open health data



Data in Brief (DiB) Dataverse (Elsevier) Home

Harvard Dataverse > Data in Brief (DiB) Dataverse > High resolution 3D laboratory x-ray tomography data of femora from young, 1-14 day old C57BL/6 mice

High resolution 3D laboratory x-ray tomography data of femora from young, 1-14 day old C57BL/6 mice

Bortel, Emely L; Duda, Georg N; Mundlos, Stefan; Willie, Bettina M; Fratzl, Peter; Zaslansky, Paul, 2015, "High resolution 3D laboratory x-ray tomography data of femora from young, 1-14 day old C57BL/6 mice", http://dx.doi.org/10.7910/DVN/29628, Harvard Dataverse, V1

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description This data article contains high resolution (1.2 μm effective pixel size) lab-based micro-computed tomography (μCT) reconstructed volume data of young C57BL/6 mouse femur bone midshafts. This data formed the basis for the analyses of bone structure development in healthy mice, including closed and open porosity as reported in http://dx.doi.org/10.1016/j.actbio.2015.03.027. The data reveal changes seen in bone material and porosity distribution in young animals aged 1 to 14 days old.The mouse bones transform from porous scaffolds into solid structures during normal organogenesis. The large data may be freely used by others and in all research areas.

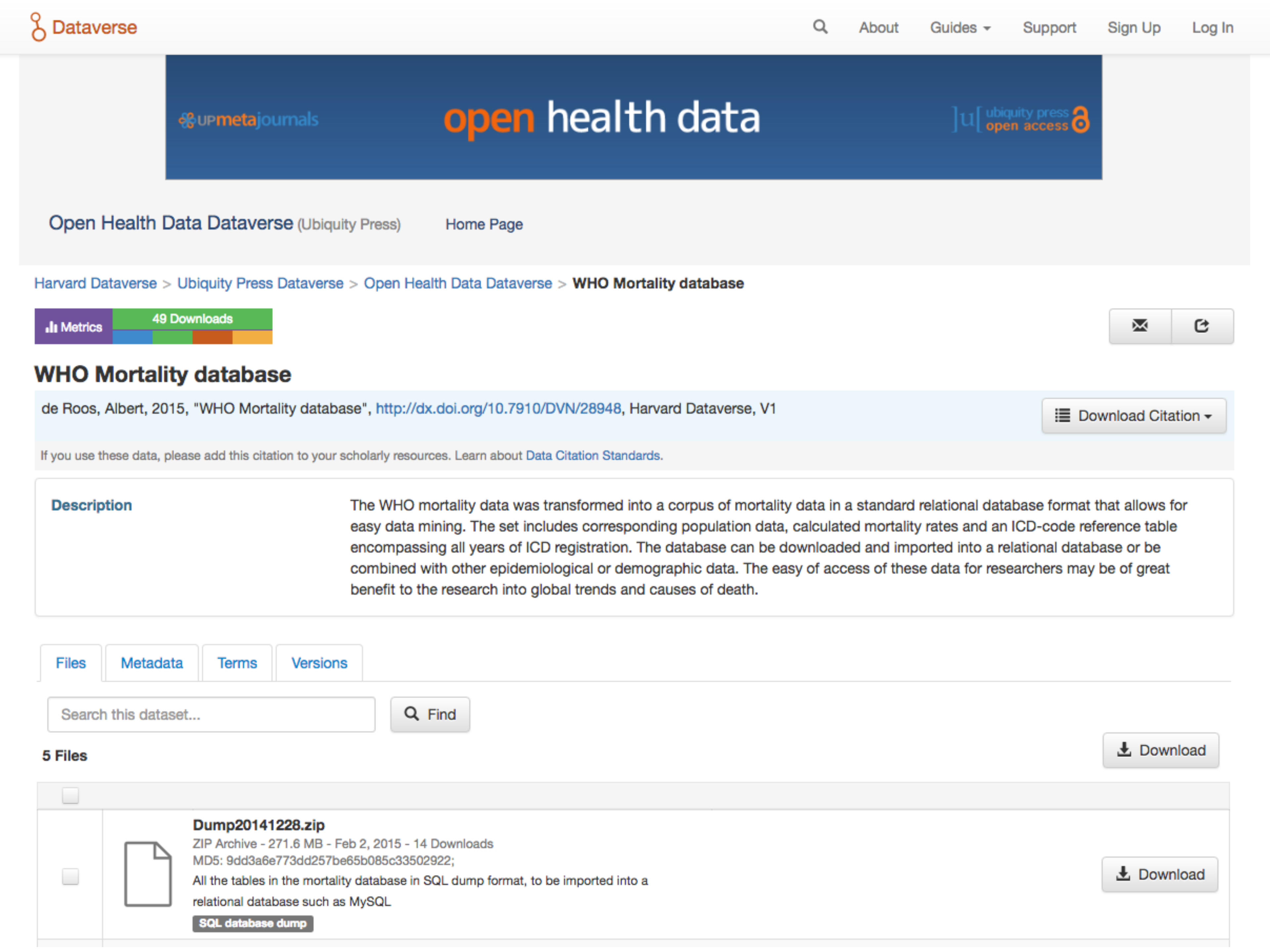
Keyword C57BL/6 growth

Related Publication Bortel EL, Duda GN, Mundlos S, Willie BM, Fratzl P, Zaslansky P. Long bone maturation is driven by pore closing: A quantitative tomography investigation of structural formation in young C57BL/6 mice. Acta Biomater (2015)

19 Files

□

□ **day10_sample1.zip**
ZIP Archive - 1.3 GB - Apr 6, 2015 - 16 Downloads
MD5: 2f03c776e509e1dac5cf212efd78fbf3;



Harvard Dataverse > Ubiquity Press Dataverse > Open Health Data Dataverse > **WHO Mortality database**

WHO Mortality database

de Roos, Albert, 2015, "WHO Mortality database", http://dx.doi.org/10.7910/DVN/28948, Harvard Dataverse, V1

If you use these data, please add this citation to your scholarly resources. Learn about Data Citation Standards.

Description The WHO mortality data was transformed into a corpus of mortality data in a standard relational database format that allows for easy data mining. The set includes corresponding population data, calculated mortality rates and an ICD-code reference table encompassing all years of ICD registration. The database can be downloaded and imported into a relational database or be combined with other epidemiological or demographic data. The easy of access of these data for researchers may be of great benefit to the research into global trends and causes of death.

5 Files

□

□ **Dump20141228.zip**
ZIP Archive - 271.6 MB - Feb 2, 2015 - 14 Downloads
MD5: 9dd3a6e773dd257be65b085c33502922;
All the tables in the mortality database in SQL dump format, to be imported into a relational database such as MySQL
SQL database dump

Future Work w/ bioCADDIE

Will be attending bioCADDIE workshop in late June to learn more from you!

Thank you!

Contact: support@dataverse.org

Twitter: @dataverseorg

Web: http://dataverse.org