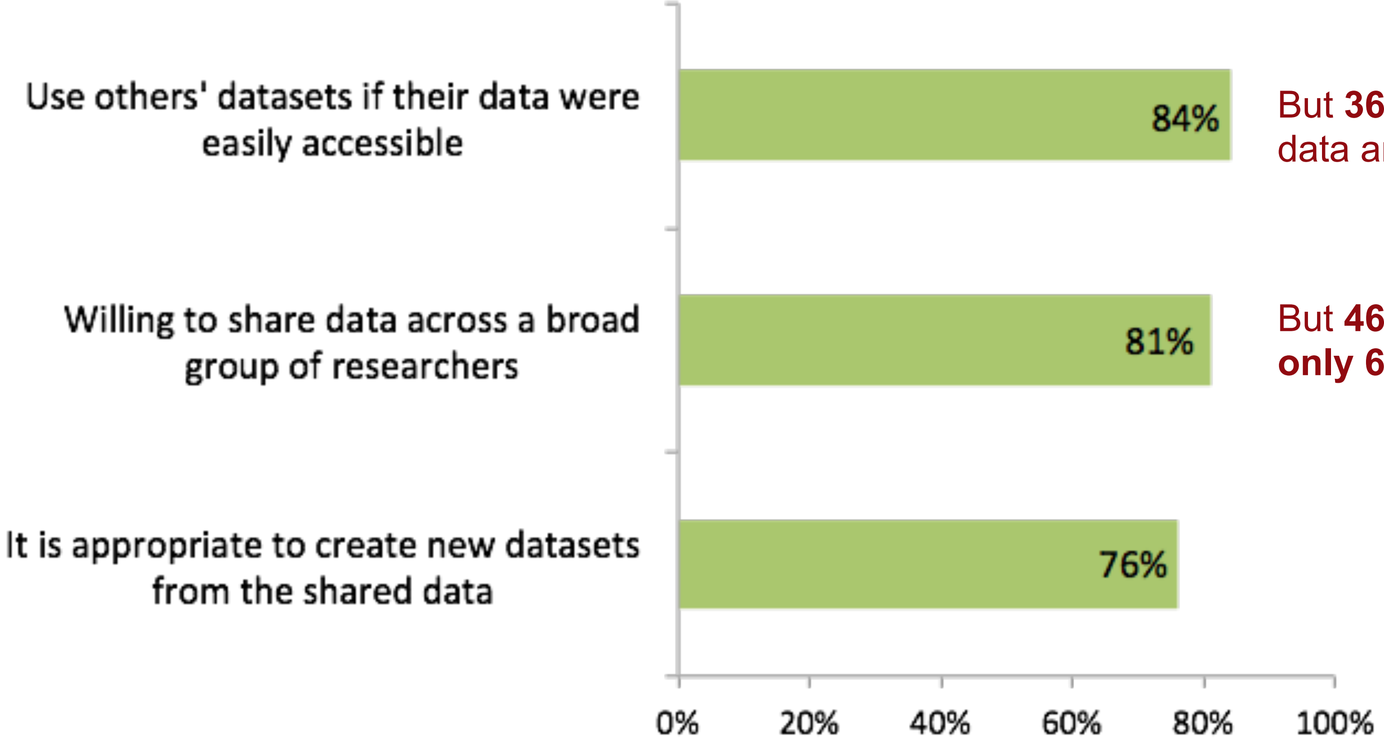
The Care and Feeding of Scientific Data

Mercè Crosas @mercecrosas
Director of Data Science, IQSS, Harvard Univeristy

On Data Sharing:
What researchers want and what researchers do

Online survey with 1315 respondents across disciplines

(9% response rate, mostly members of DataONE):



But **36%** agreed that their own data are easy to access

But **46%** shared some data, and **only 6%** shared all data

Tenopir, Allard, Douglass, Aydinoglu, Wu, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101 (Figure acknowledgement: Tenopir, U. of Tennessee)

Researchers intent vs researchers actions

Ten-year study with 22 random participants from the Center for Embedded Network Sensing (CENS):

“Data sharing tends to occur only through interpersonal exchanges.”

“10 of the 22 participants were unaware of repositories that would accept data from their type of research.”

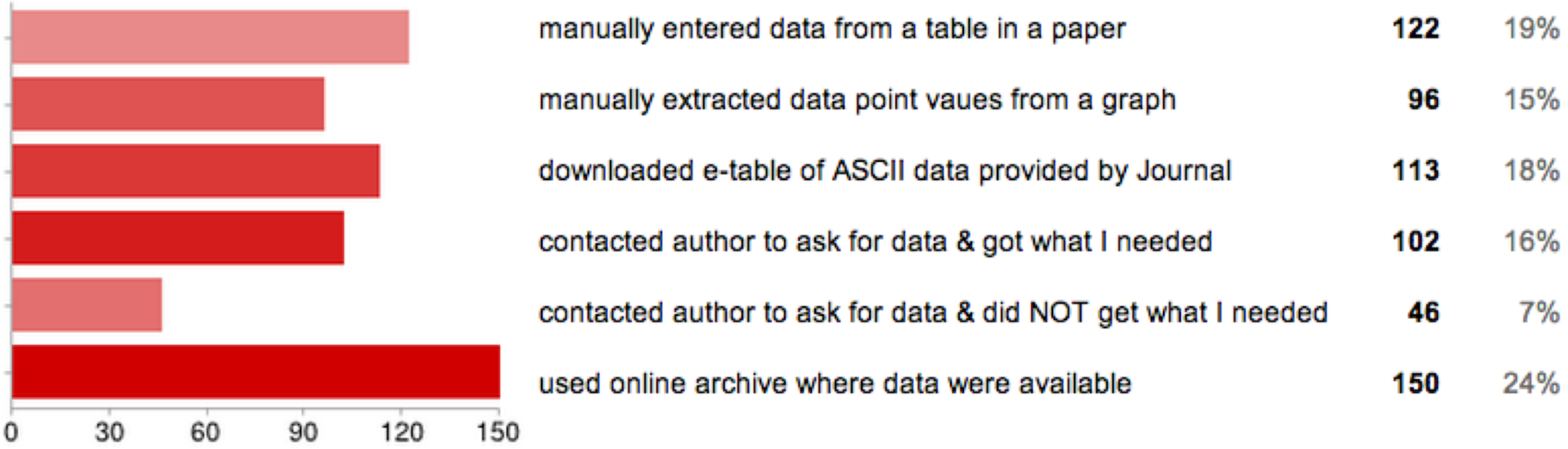
“14 participants said that they use data they themselves did not generate”

Wallis JC, Rolando E, Borgman CL (2013) If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8(7): e67332. doi:10.1371/journal.pone.0067332

Data sharing is mostly demand-driven

Survey sent to ~ 350 researchers at the Harvard-Smithsonian Center for Astrophysics; 175 respondents:

Have you ever used DATA you learned about from reading a Journal article?

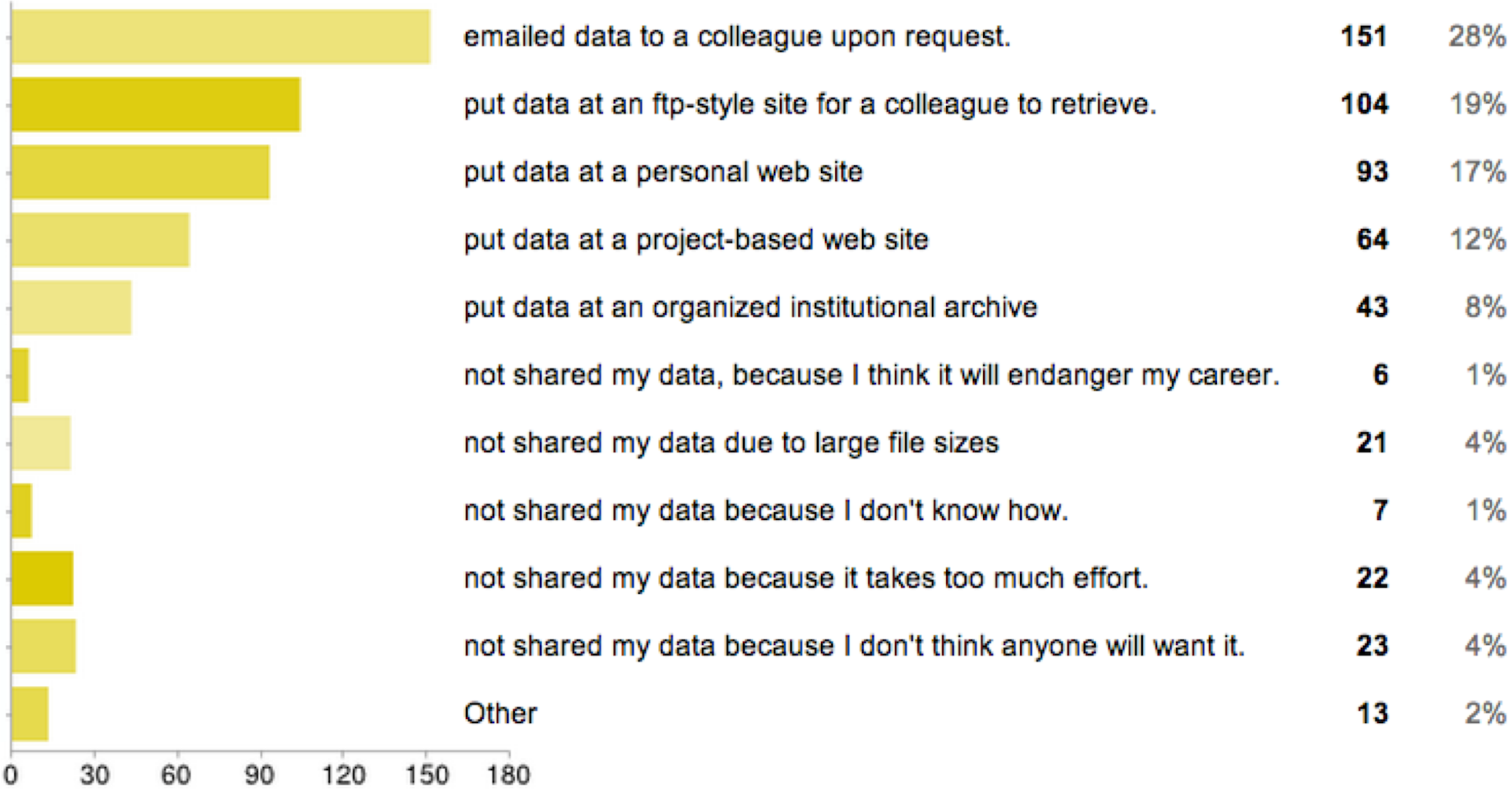


manually entered data from a table in a paper	**122**	19%
manually extracted data point vaues from a graph	**96**	15%
downloaded e-table of ASCII data provided by Journal	**113**	18%
contacted author to ask for data & got what I needed	**102**	16%
contacted author to ask for data & did NOT get what I needed	**46**	7%
used online archive where data were available	**150**	24%

Pepe, Goodman, Muench, Crosas, Erdmann (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 9 (8): e104798. doi:10.1371/journal.pone.0104798

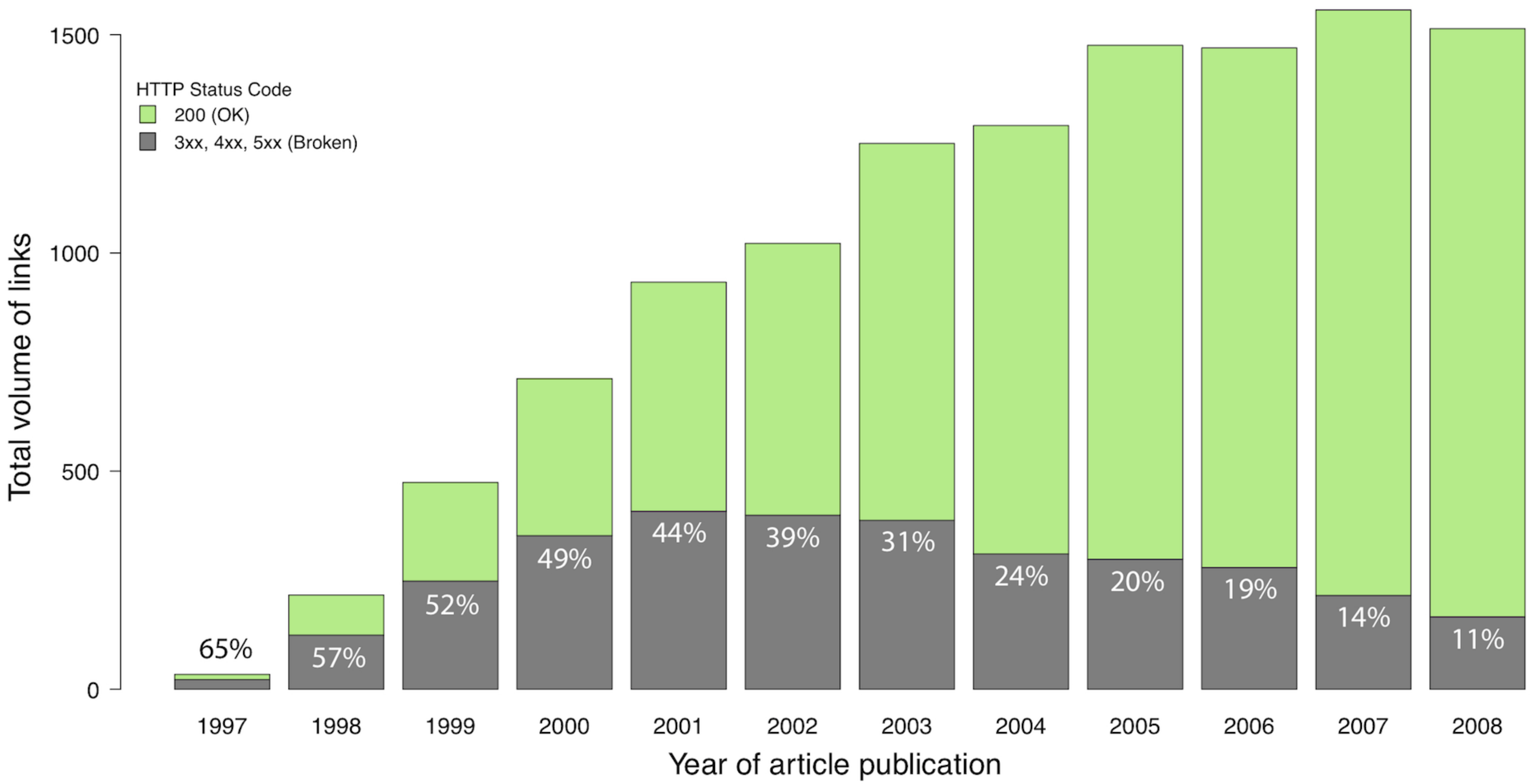
Data are accessed in various ways for reuse

When it comes to sharing DATA you've created, collected or curated, you have?

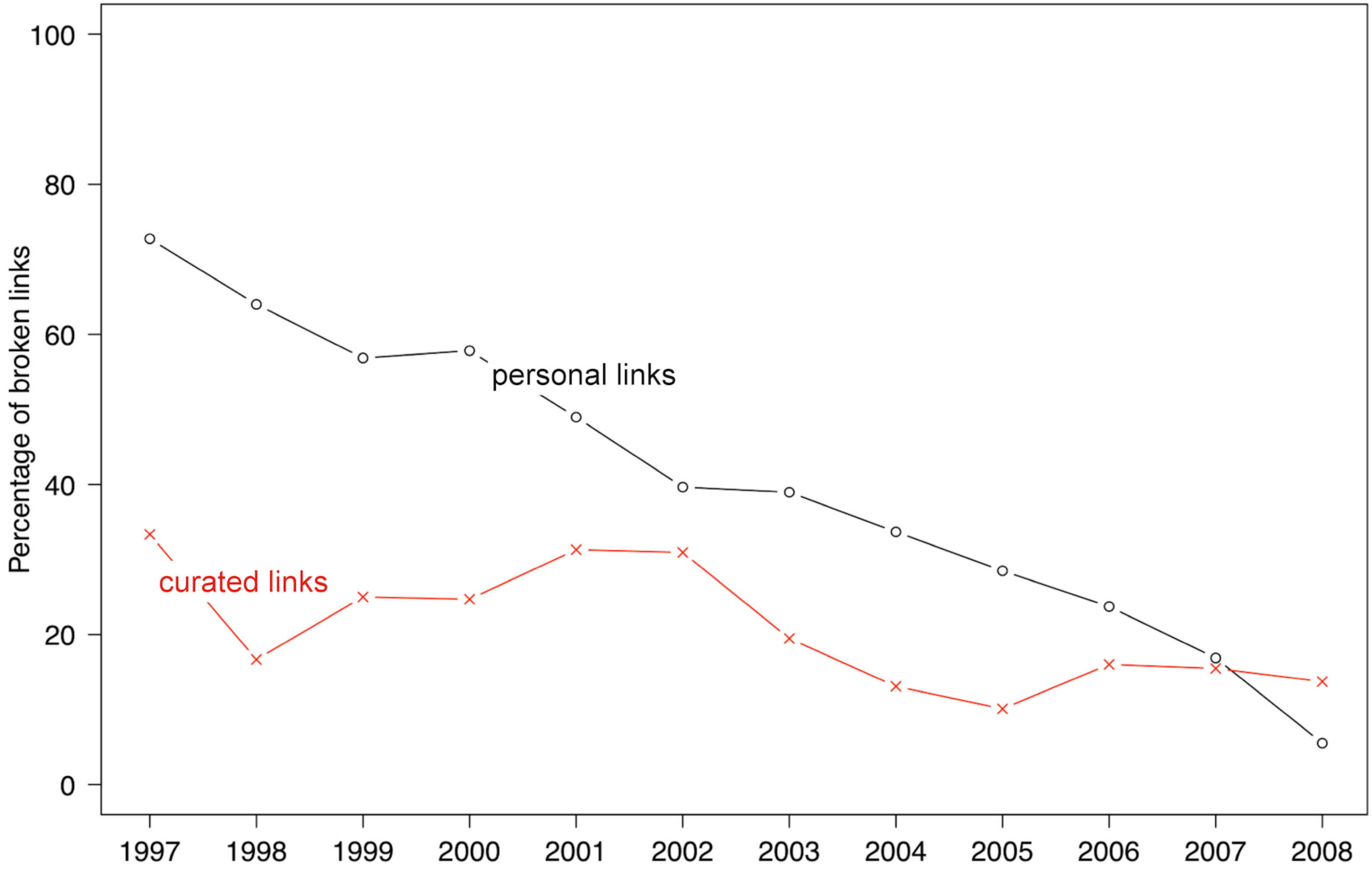


emailed data to a colleague upon request.	151	28%
put data at an ftp-style site for a colleague to retrieve.	104	19%
put data at a personal web site	93	17%
put data at a project-based web site	64	12%
put data at an organized institutional archive	43	8%
not shared my data, because I think it will endanger my career.	6	1%
not shared my data due to large file sizes	21	4%
not shared my data because I don't know how.	7	1%
not shared my data because it takes too much effort.	22	4%
not shared my data because I don't think anyone will want it.	23	4%
Other	13	2%

I'll share my data when you ask me



Links to data from 4 astronomy journals over 10 yrs



After 10 yrs since publication, >70% broken links

We can do better

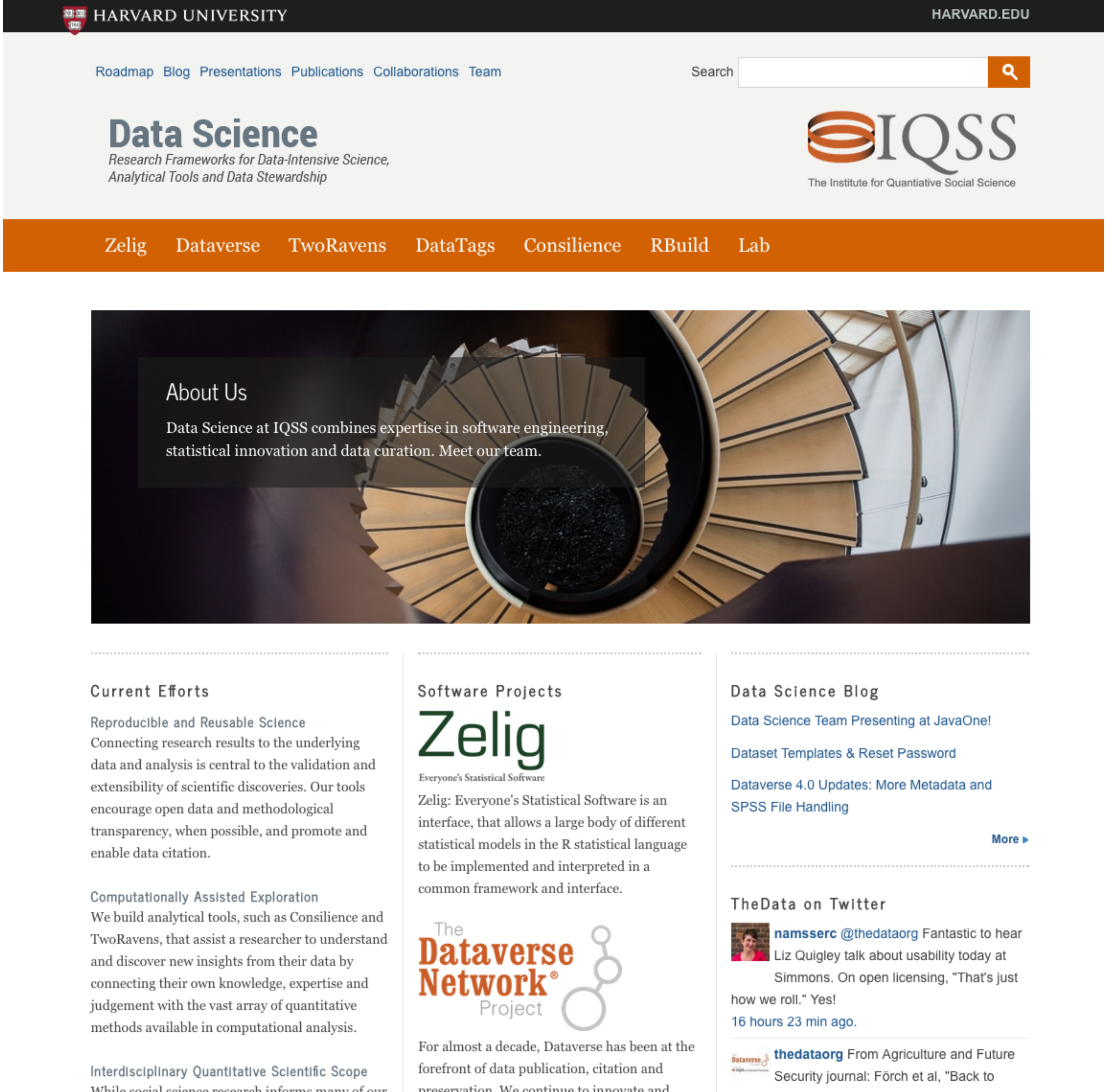
10 Simple Rules

1. Love your data, and let others love it too
2. Share your data online, with a permanent identifier
3. Conduct science with data reuse in mind
4. Publish workflow as context
5. Link your data to your publications as early as possible
6. Publish your code
7. Say how you want to get credit for your data
8. Foster and use data repositories
9. Reward colleagues who share their data properly
10. Help establish data science and data scientist as vital

Goodman, Pepe, Blocker, Borgman, Cranmer, Crosas, Di Stefano, Gil, Groth, Hogg, Kashyap, Hedstrom, Mahabal, Siemiginowska, Slavkovic (2014), 10 Simple Rules for the Care and Feeding of Scientific Data, PLOS Computational Biology

A two-pronged approach to motivate cultural and policy change:

- Engage in policy debate, participate in community initiatives, and write papers like the "10 Simple Rules"
- Provide technical solutions to facilitate data sharing, reusability and interoperability



HARVARD UNIVERSITY HARVARD.EDU

Roadmap Blog Presentations Publications Collaborations Team

Search

Data Science

Research Frameworks for Data-Intensive Science, Analytical Tools and Data Stewardship

IQSS
The Institute for Quantiative Social Science

Zelig Dataverse TwoRavens DataTags Consilience RBuild Lab

About Us

Data Science at IQSS combines expertise in software engineering, statistical innovation and data curation. Meet our team.

Current Efforts

Reproducible and Reusable Science

Connecting research results to the underlying data and analysis is central to the validation and extensibility of scientific discoveries. Our tools encourage open data and methodological transparency, when possible, and promote and enable data citation.

Computationally Assisted Exploration

We build analytical tools, such as Consilience and TwoRavens, that assist a researcher to understand and discover new insights from their data by connecting their own knowledge, expertise and judgement with the vast array of quantitative methods available in computational analysis.

Interdisciplinary Quantitative Scientific Scope

While social science research informs many of our

Software Projects

Zelig
Everyone's Statistical Software

Zelig: Everyone's Statistical Software is an interface, that allows a large body of different statistical models in the R statistical language to be implemented and interpreted in a common framework and interface.

The Dataverse Network® Project

For almost a decade, Dataverse has been at the forefront of data publication, citation and preservation. We continue to innovate and

Data Science Blog

Data Science Team Presenting at JavaOne!

Dataset Templates & Reset Password

Dataverse 4.0 Updates: More Metadata and SPSS File Handling

More ▸

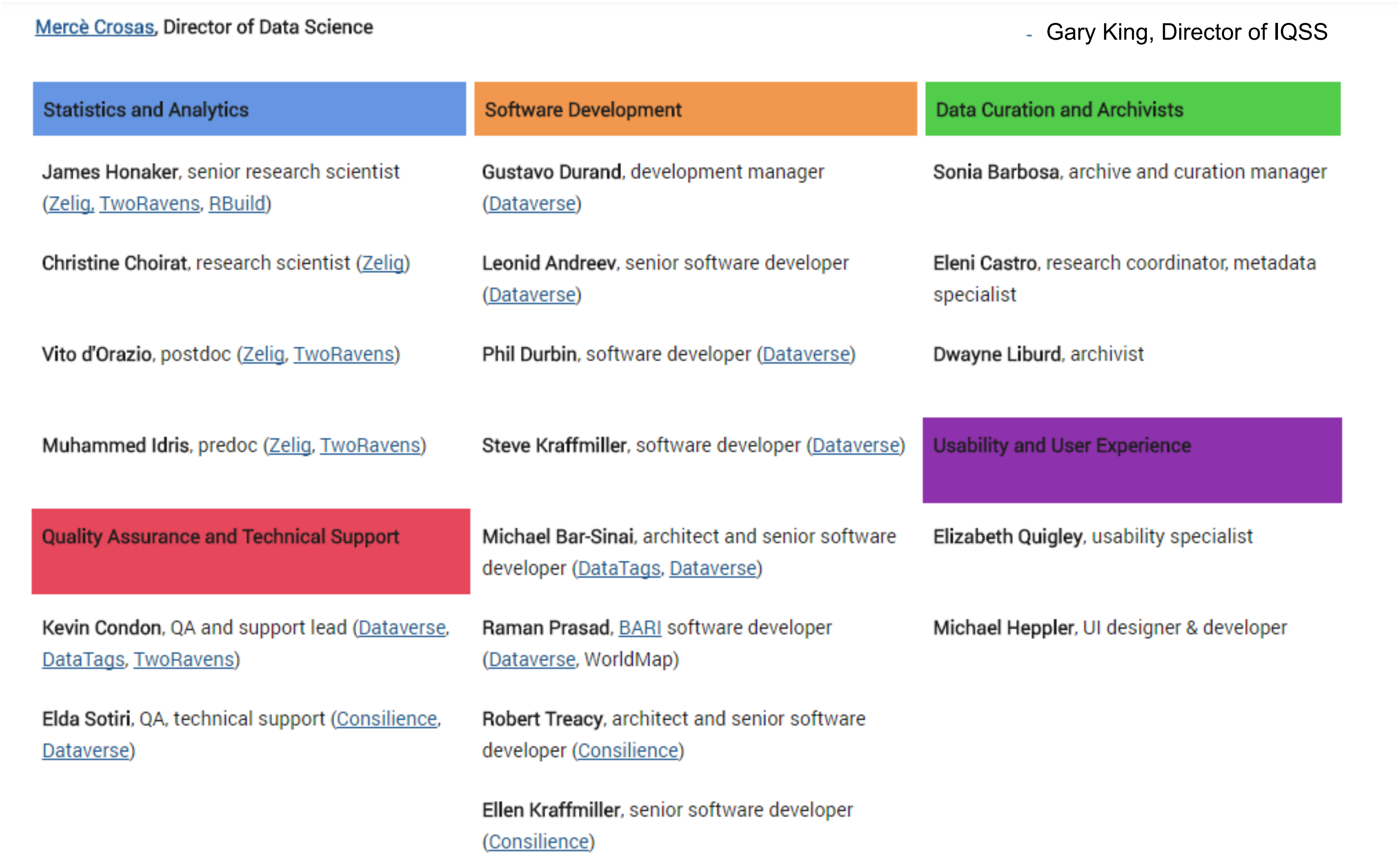
TheData on Twitter

namsserc @thedataorg Fantastic to hear Liz Quigley talk about usability today at Simmons. On open licensing, "That's just how we roll." Yes!
16 hours 23 min ago.

thedataorg From Agriculture and Future Security journal: Förch et al, "Back to

IQSS Data Science Team members

- Gary King, Director of IQSS



Mercè Crosas, Director of Data Science

Statistics and Analytics

James Honaker, senior research scientist (Zelig, TwoRavens, RBuild)

Christine Choirat, research scientist (Zelig)

Vito d'Orazio, postdoc (Zelig, TwoRavens)

Muhammed Idris, predoc (Zelig, TwoRavens)

Quality Assurance and Technical Support

Kevin Condon, QA and support lead (Dataverse, DataTags, TwoRavens)

Elda Sotiri, QA, technical support (Consilience, Dataverse)

Software Development

Gustavo Durand, development manager (Dataverse)

Leonid Andreev, senior software developer (Dataverse)

Phil Durbin, software developer (Dataverse)

Steve Kraffmiller, software developer (Dataverse)

Michael Bar-Sinai, architect and senior software developer (DataTags, Dataverse)

Raman Prasad, BARI software developer (Dataverse, WorldMap)

Robert Treacy, architect and senior software developer (Consilience)

Ellen Kraffmiller, senior software developer (Consilience)

Data Curation and Archivists

Sonia Barbosa, archive and curation manager

Eleni Castro, research coordinator, metadata specialist

Dwayne Liburd, archivist

Usability and User Experience

Elizabeth Quigley, usability specialist

Michael Heppler, UI designer & developer

Dataverse: A bridge between traditional archives and posting data in your website

Traditional data archives

Professional curation
Full preservation

Posting data on the web

No curation or
preservation guaranteed

The
Dataverse
Network®
Project

Infrastructure
to curate and
preserve data

Persistence guaranteed
by hosting institution

control and
credit for data
author

Tools to facilitate curation
and preservation

Dataverse Community

Federated Dataverses around the world with **persistence guaranteed by**:

VE RI TAS HARVARD

UNC ODUM

Data Archiving and Networked Services DANS

OCUL Ontario Council of University Libraries

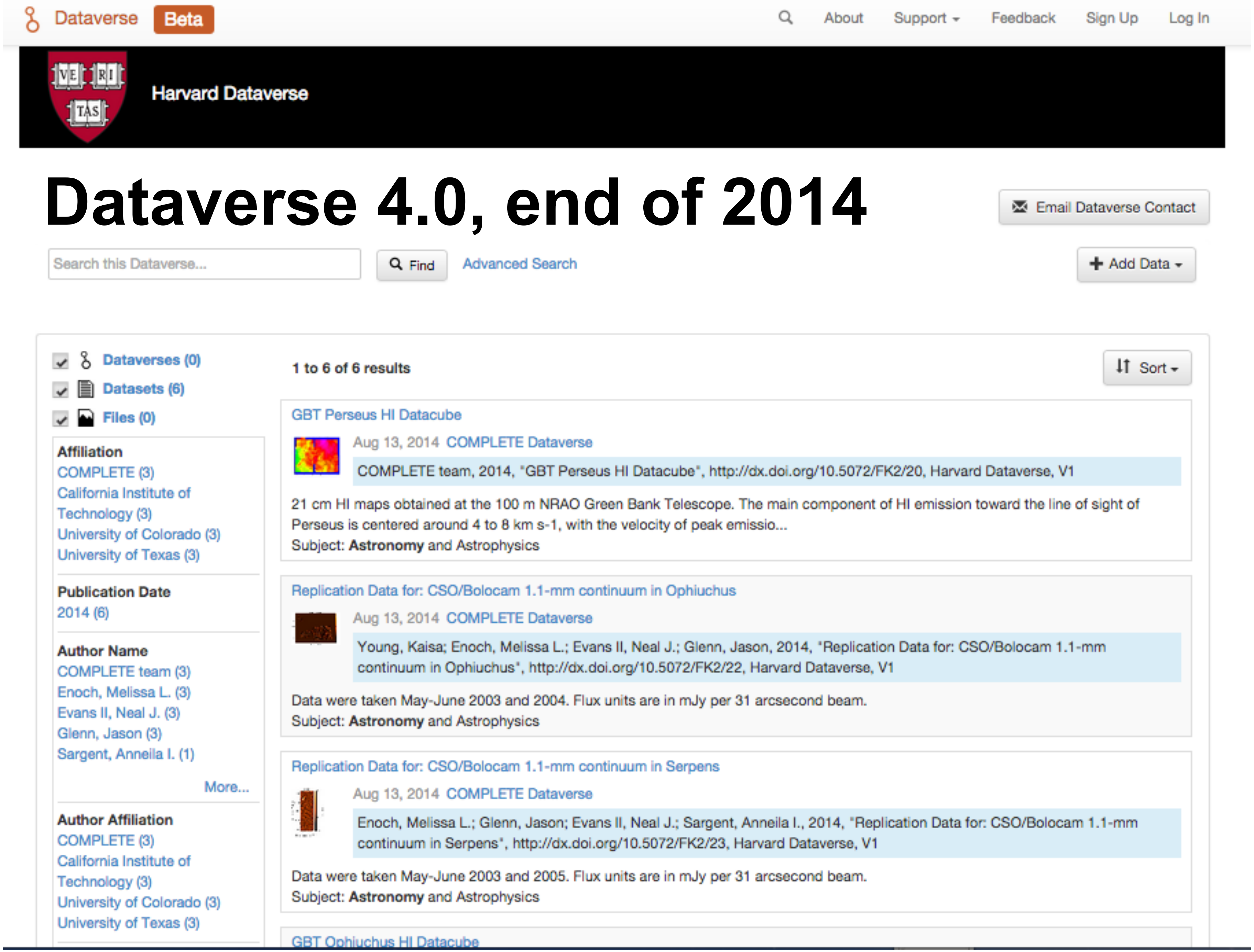
FUDAN UNIVERSITY 1905

...

- Dataverse.org coming at the end of 2014
- Dataverse advisory team and community groups:
    - API: common repository deposit API; search and data API
    - Metadata: standards per domain; automate extraction
    - Storage: multiple storages; integrate with iRODS
    - Preservation: integrate with archival and preservation tools
    - Authentication: multiple identity providers
    - Internationalization: chinese, spanish

Upcoming software improvements and new features

Dataverse 4.0, end of 2014



☑ Dataverses (0)
☑ Datasets (6)
☑ Files (0)

Affiliation
COMPLETE (3)
California Institute of Technology (3)
University of Colorado (3)
University of Texas (3)

Publication Date
2014 (6)

Author Name
COMPLETE team (3)
Enoch, Melissa L. (3)
Evans II, Neal J. (3)
Glenn, Jason (3)
Sargent, Anneila I. (1)
More...

Author Affiliation
COMPLETE (3)
California Institute of Technology (3)
University of Colorado (3)
University of Texas (3)

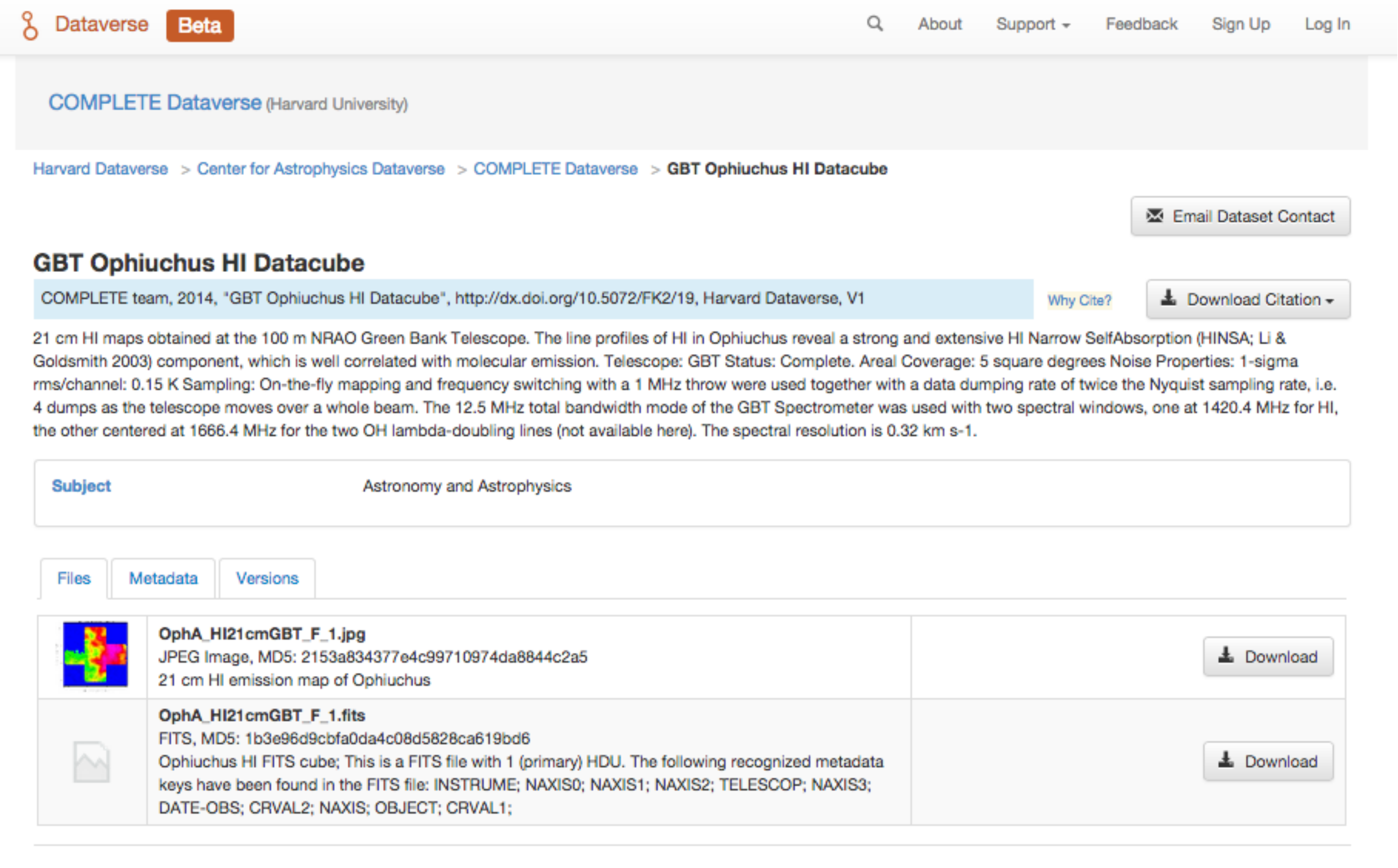
1 to 6 of 6 results

GBT Perseus HI Datacube
Aug 13, 2014 COMPLETE Dataverse
COMPLETE team, 2014, "GBT Perseus HI Datacube", http://dx.doi.org/10.5072/FK2/20, Harvard Dataverse, V1
21 cm HI maps obtained at the 100 m NRAO Green Bank Telescope. The main component of HI emission toward the line of sight of Perseus is centered around 4 to 8 km s-1, with the velocity of peak emissio...
Subject: **Astronomy** and Astrophysics

Replication Data for: CSO/Bolocam 1.1-mm continuum in Ophiuchus
Aug 13, 2014 COMPLETE Dataverse
Young, Kaisa; Enoch, Melissa L.; Evans II, Neal J.; Glenn, Jason, 2014, "Replication Data for: CSO/Bolocam 1.1-mm continuum in Ophiuchus", http://dx.doi.org/10.5072/FK2/22, Harvard Dataverse, V1
Data were taken May-June 2003 and 2004. Flux units are in mJy per 31 arcsecond beam.
Subject: **Astronomy** and Astrophysics

Replication Data for: CSO/Bolocam 1.1-mm continuum in Serpens
Aug 13, 2014 COMPLETE Dataverse
Enoch, Melissa L.; Glenn, Jason; Evans II, Neal J.; Sargent, Anneila I., 2014, "Replication Data for: CSO/Bolocam 1.1-mm continuum in Serpens", http://dx.doi.org/10.5072/FK2/23, Harvard Dataverse, V1
Data were taken May-June 2003 and 2005. Flux units are in mJy per 31 arcsecond beam.
Subject: **Astronomy** and Astrophysics

GBT Ophiuchus HI Datacube



COMPLETE Dataverse (Harvard University)

Harvard Dataverse > Center for Astrophysics Dataverse > COMPLETE Dataverse > **GBT Ophiuchus HI Datacube**

GBT Ophiuchus HI Datacube

COMPLETE team, 2014, "GBT Ophiuchus HI Datacube", http://dx.doi.org/10.5072/FK2/19, Harvard Dataverse, V1

21 cm HI maps obtained at the 100 m NRAO Green Bank Telescope. The line profiles of HI in Ophiuchus reveal a strong and extensive HI Narrow SelfAbsorption (HINSA; Li & Goldsmith 2003) component, which is well correlated with molecular emission. Telescope: GBT Status: Complete. Areal Coverage: 5 square degrees Noise Properties: 1-sigma rms/channel: 0.15 K Sampling: On-the-fly mapping and frequency switching with a 1 MHz throw were used together with a data dumping rate of twice the Nyquist sampling rate, i.e. 4 dumps as the telescope moves over a whole beam. The 12.5 MHz total bandwidth mode of the GBT Spectrometer was used with two spectral windows, one at 1420.4 MHz for HI, the other centered at 1666.4 MHz for the two OH lambda-doubling lines (not available here). The spectral resolution is 0.32 km s-1.

Subject	Astronomy and Astrophysics

OphA_HI21cmGBT_F_1.jpg
JPEG Image, MD5: 2153a834377e4c99710974da8844c2a5
21 cm HI emission map of Ophiuchus

OphA_HI21cmGBT_F_1.fits
FITS, MD5: 1b3e96d9cbfa0da4c08d5828ca619bd6
Ophiuchus HI FITS cube; This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file: INSTRUME; NAXIS0; NAXIS1; NAXIS2; TELESCOP; NAXIS3; DATE-OBS; CRVAL2; NAXIS; OBJECT; CRVAL1;

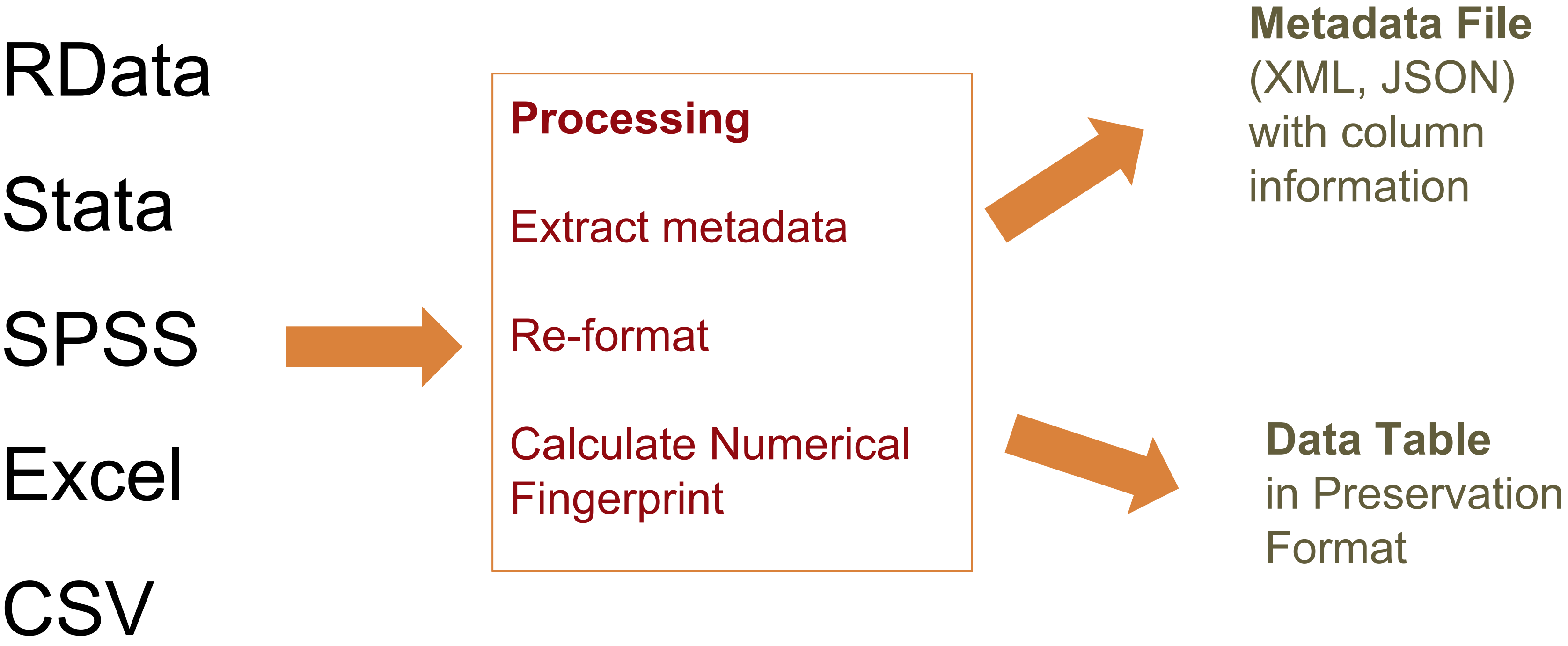
A Dataset may contain any type of files, including code

Extensive Metadata, with data reuse in mind

- Descriptive metadata
  - **Citation Metadata** for all (compliant with DataCite)
  - **Domain metadata** blocks:
    - Social Sciences (compliant with DDI)
    - Biomedical (compliant with ISA-Tab)
    - Astronomy (compliant with VO)
    - Custom
- File Level metadata
  - **Automated extraction** of variables/columns

    metadata from R data, Stata, SPSS, Excel, CSV, and header metadata from FITS

Automated Data Processing



RData

Stata

SPSS

Excel

CSV

Processing

Extract metadata

Re-format

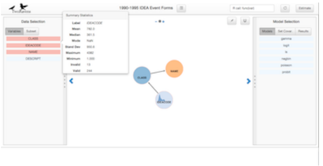
Calculate Numerical Fingerprint

Metadata File
(XML, JSON)
with column
information

Data Table
in Preservation
Format

Data Exploration and Analysis Tools

Tabular data

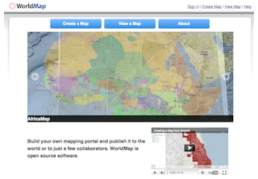


Summary Statistics

Label	IDEACODE
Mean	782.0
Median	361.5
Mode	NaN
Stand Dev	900.6
Maximum	4382
Minimum	1.000
Invalid	13
Valid	244

TwoRavens:
Statistical analysis

Data with geo-references


WorldMap
Sign in | Create Map | View Map | Help
Create a Map
View a Map
About
AfricaMap
Creating a Map from Scratch

Build your own mapping portal and publish it to the world or to just a few collaborators. WorldMap is open source software.

WorldMap:
Statistical analysis

Survey data

Survey Tool:
cross-tabulations and reports

Data with time variable

Time-series Visualizations:
explore time series data

Open Licenses and Terms of Use

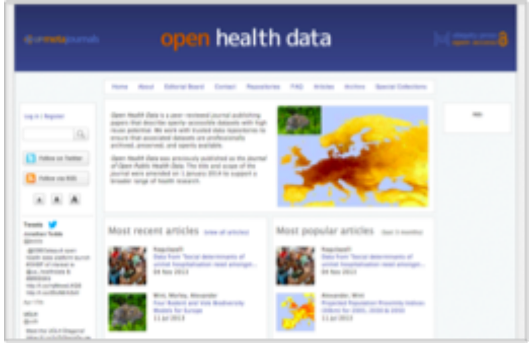
Multiple levels of access and reuse:

- Open License (CC0), with an understanding that scientific communication is based on attribution
- Custom Terms of Use
- Metadata open and files restricted: access may be granted upon request

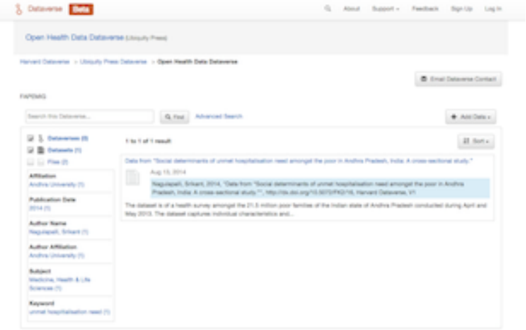
On going collaborations

Automated Data Publishing

Journal Publishing System



Journal Dataverse



Open Health Data Dataverse (Ubiquity Press)

Harvard Dataverse > Ubiquity Press Dataverse > Open Health Data Dataverse

1 to 1 of 1 result

Data from "Social determinants of unmet hospitalisation need amongst the poor in Andhra Pradesh, India: A cross-sectional study."

Aug 13, 2014

The dataset is of a health survey amongst the 21.5 million poor families of the Indian state of Andhra Pradesh conducted during April and May 2013. The dataset captures individual characteristics and...

Integration of publishing systems with data repositories via API

Towards a **common API** across repositories and publishing systems

PKP
PUBLIC KNOWLEDGE PROJECT

The Institute *for* Quantitative Social Science at Harvard University

ALFRED P. SLOAN FOUNDATION

DataBridge

DATABRIDGE

- Connect data to data (by analyzing metadata and usage)
- Connect data to users (via ORCID)

UNC

renci
RESEARCH ENGAGEMENT INNOVATION

The Institute
for Quantitative Social Science
at Harvard University

NSF

Data Citation and Provenance

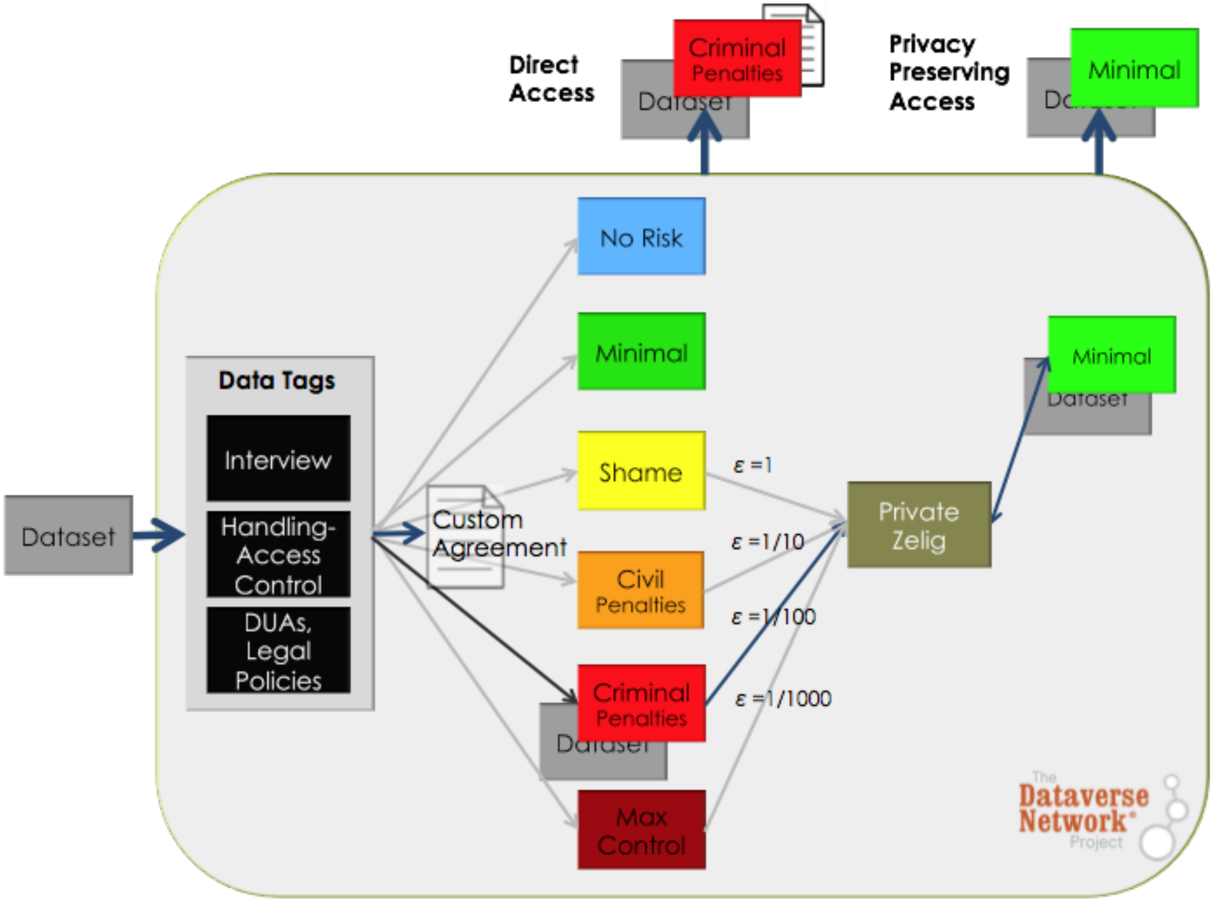
- Incorporate provenance in data citation:
  - As metadata
  - DOI to provenance object
- Tracking multiple transformations:
  - disclosed provenance (e.g., explicit SQL query)
  - observed provenance (e.g., functions executed in R)

HARVARD
School of Engineering
and Applied Sciences

The Institute
for Quantitative Social Science
at Harvard University

NSF

Sharing Sensitive Data



DataTags

HARVARD School of Engineering and Applied Sciences

The Institute *for* Quantitative Social Science at Harvard University

Berkman The Berkman Center for Internet & Society at Harvard University

NSF

Thank you

@mercecrosas