Sharing Data with Dataverse

Mercè Crosas, Ph.D.
Director of Product Development
Institute for Quantitative Social Science (IQSS)
Harvard University

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

1. Data Sharing and Replication

2. A Solution with the Dataverse Network

SHARE your data

it's good for you, and for the world.

Come. Eat lunch. Accelerate the pace of science.

CfA, PHILLIPS Auditorium, 11:45 MONDAY 4/2/12

theastrodata.org

POWERED BY THE Dataverse Network™ PROJECT & SEAMLESS ASTRONOMY Linking scientific data, publications, and communities

Slide acknowledgment: Alyssa Goodman

From Data Sharing to Replication

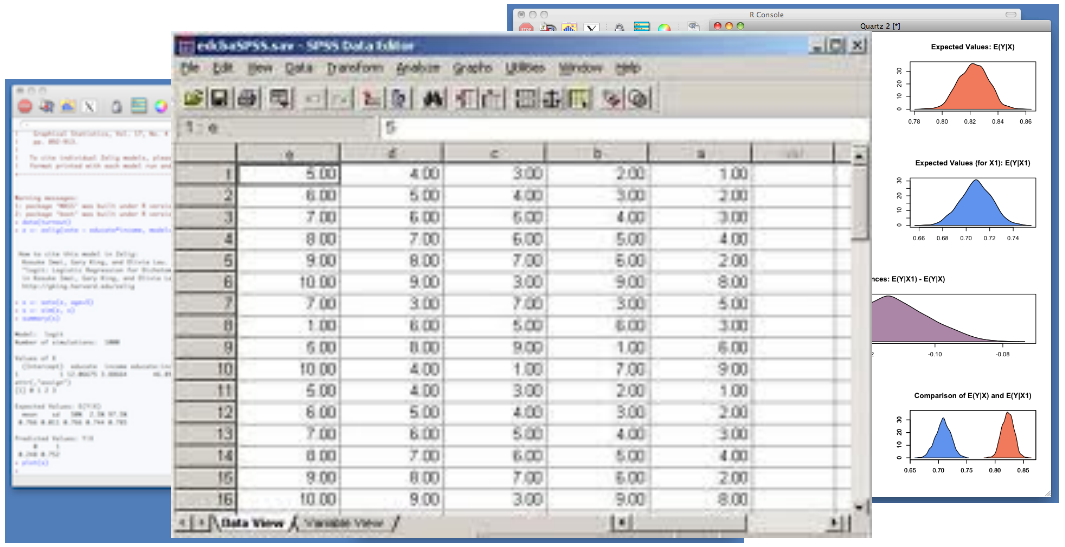
"The replication standard holds that **sufficient information** exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author."

King, G. 1995 "Replication, Replication"

Published Work

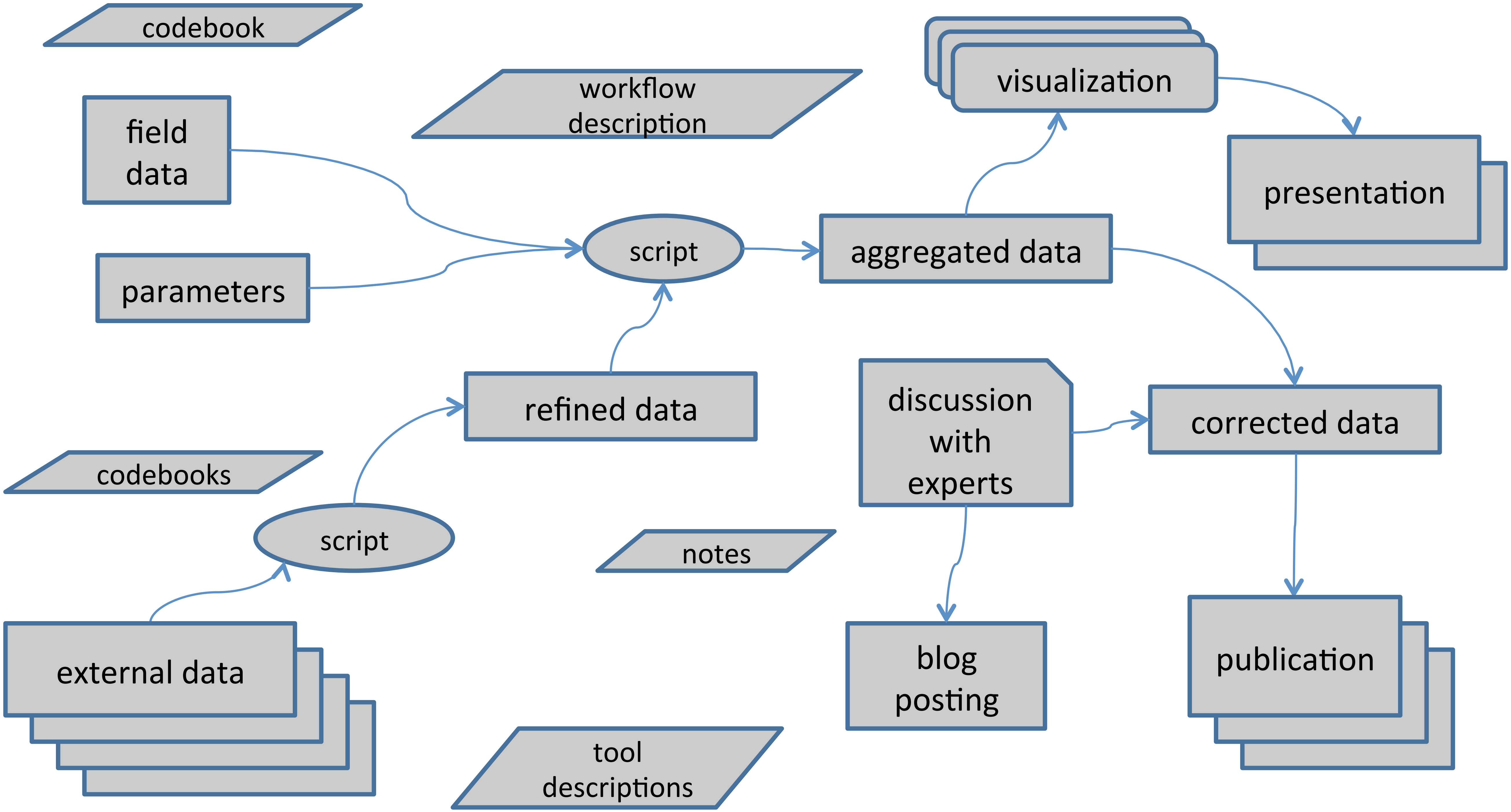
+

Data + Metadata + Supporting Files


R Console
Quartz 2 [*]
edk3aSPSS.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : e
5
Data View
Variable View
Expected Values: E(Y|X)
Expected Values (for X1): E(Y|X1)
nces: E(Y|X1) - E(Y|X)
Comparison of E(Y|X) and E(Y|X1)

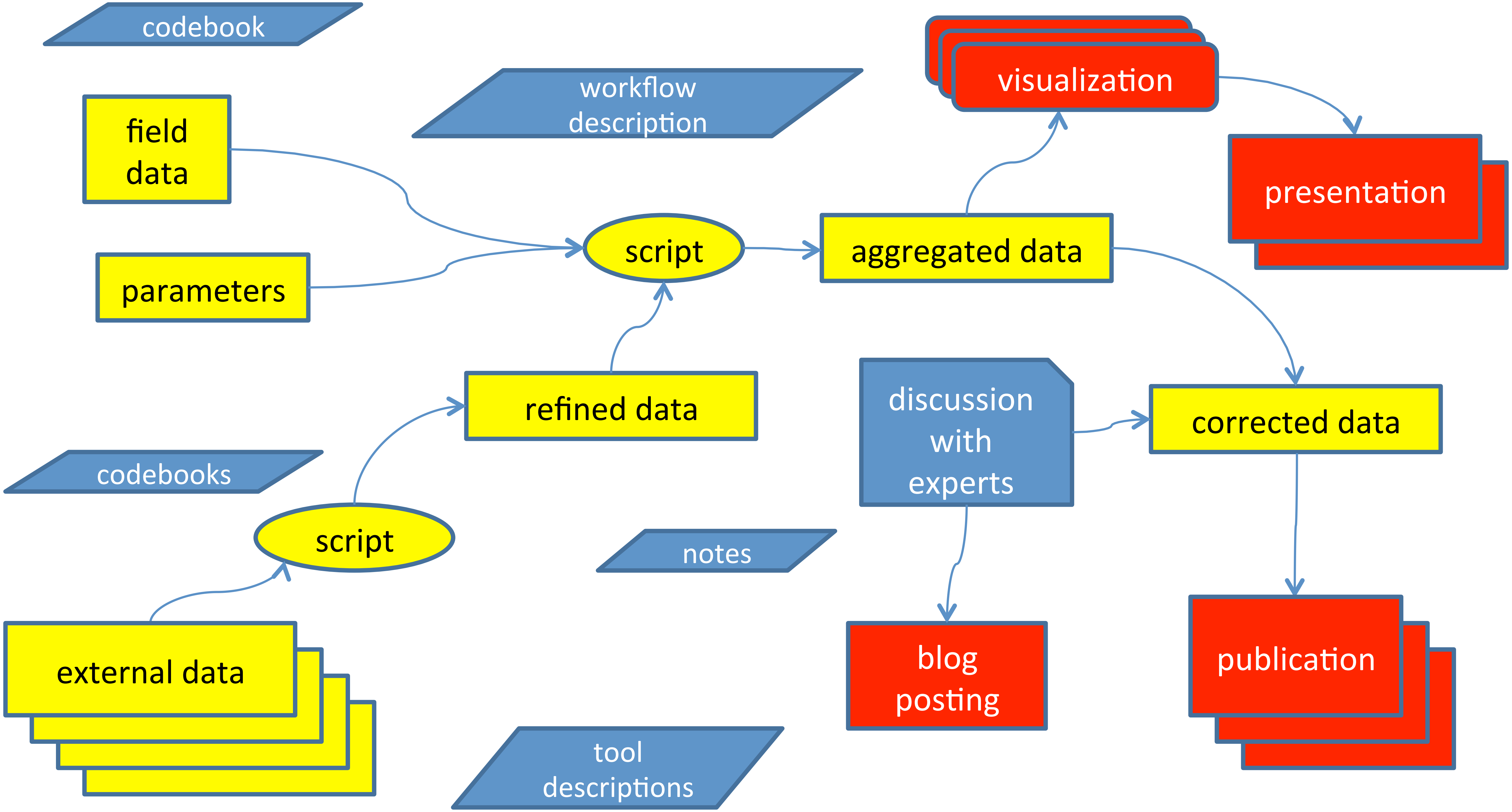
	e	d	c	b	a	var
1	5.00	4.00	3.00	2.00	1.00	
2	6.00	5.00	4.00	3.00	2.00	
3	7.00	6.00	5.00	4.00	3.00	
4	8.00	7.00	6.00	5.00	4.00	
5	9.00	8.00	7.00	6.00	2.00	
6	10.00	9.00	3.00	9.00	8.00	
7	7.00	3.00	7.00	3.00	5.00	
8	1.00	6.00	5.00	6.00	3.00	
9	6.00	8.00	9.00	1.00	6.00	
10	10.00	4.00	1.00	7.00	9.00	
11	5.00	4.00	3.00	2.00	1.00	
12	6.00	5.00	4.00	3.00	2.00	
13	7.00	6.00	5.00	4.00	3.00	
14	8.00	7.00	6.00	5.00	4.00	
15	9.00	8.00	7.00	6.00	2.00	
16	10.00	9.00	3.00	9.00	8.00	

"Sufficient Information"?



Slide acknowledgment: Andrea Goethals

“Sufficient Information”?



codebook

workflow description

visualization

field data

presentation

script

aggregated data

parameters

refined data

discussion with experts

corrected data

codebooks

script

notes

blog posting

publication

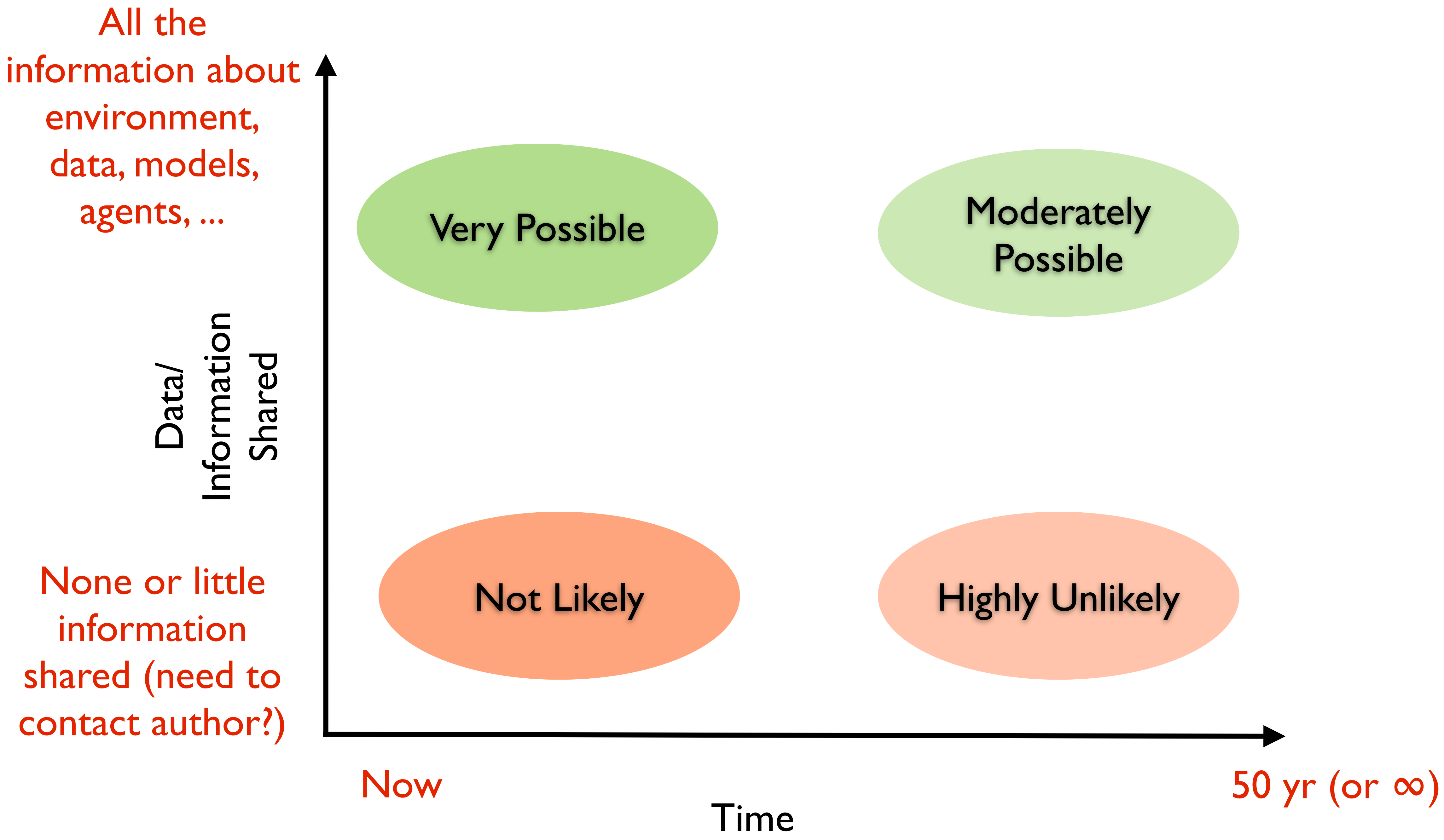
external data

tool descriptions

🟨 = data 🟦 = context (documentation) 🟥 = research products

Slide acknowledgment: Andrea Goethals

Likelihood to Replicate



All the information about environment, data, models, agents, …

Data/ Information Shared

None or little information shared (need to contact author?)

Very Possible

Moderately Possible

Not Likely

Highly Unlikely

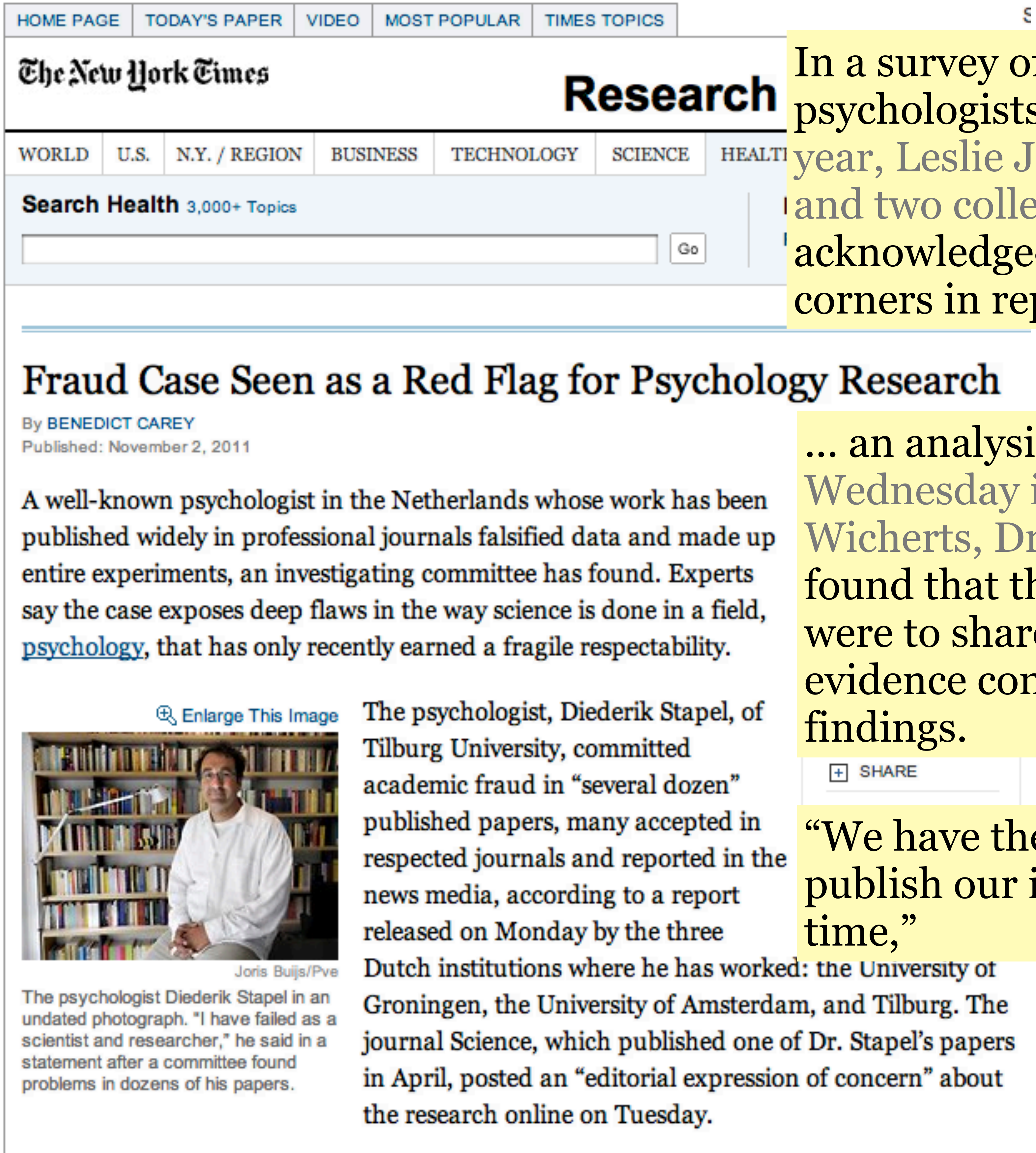
Now

Time

50 yr (or ∞)

At the very least, need sufficient documentation to judge the veracity and usefulness of the data

What can happen if we don't share



Fraud Case Seen as a Red Flag for Psychology Research

By BENEDICT CAREY
Published: November 2, 2011

A well-known psychologist in the Netherlands whose work has been published widely in professional journals falsified data and made up entire experiments, an investigating committee has found. Experts say the case exposes deep flaws in the way science is done in a field, psychology, that has only recently earned a fragile respectability.

Joris Buijs/Pve

The psychologist Diederik Stapel in an undated photograph. "I have failed as a scientist and researcher," he said in a statement after a committee found problems in dozens of his papers.

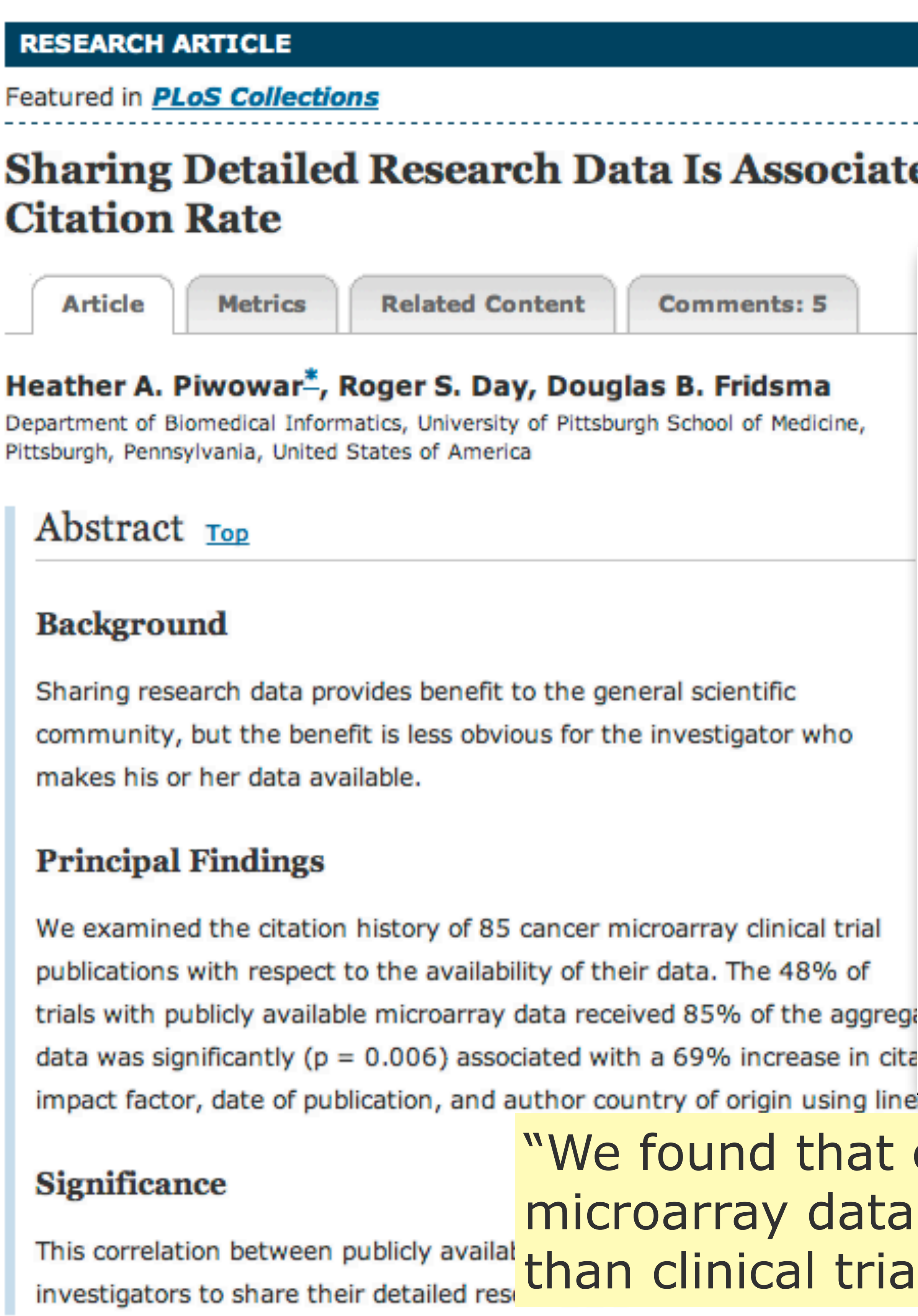
The psychologist, Diederik Stapel, of Tilburg University, committed academic fraud in "several dozen" published papers, many accepted in respected journals and reported in the news media, according to a report released on Monday by the three Dutch institutions where he has worked: the University of Groningen, the University of Amsterdam, and Tilburg. The journal Science, which published one of Dr. Stapel's papers in April, posted an "editorial expression of concern" about the research online on Tuesday.

In a survey of more than 2,000 American psychologists scheduled to be published this year, Leslie John of Harvard Business School and two colleagues found that 70 percent had acknowledged, anonymously, to cutting some corners in reporting data.

… an analysis of 49 studies appearing Wednesday in the journal PLoS One, by Dr. Wicherts, Dr. Bakker and Dylan Molenaar, found that the more reluctant that scientists were to share their data, the more likely that evidence contradicted their reported findings.

"We have the technology to share data and publish our initial hypotheses, and now's the time,"

What can happen if we share



RESEARCH ARTICLE OPEN ACCESS

Featured in PLoS Collections

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma
Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

Abstract Top

Background

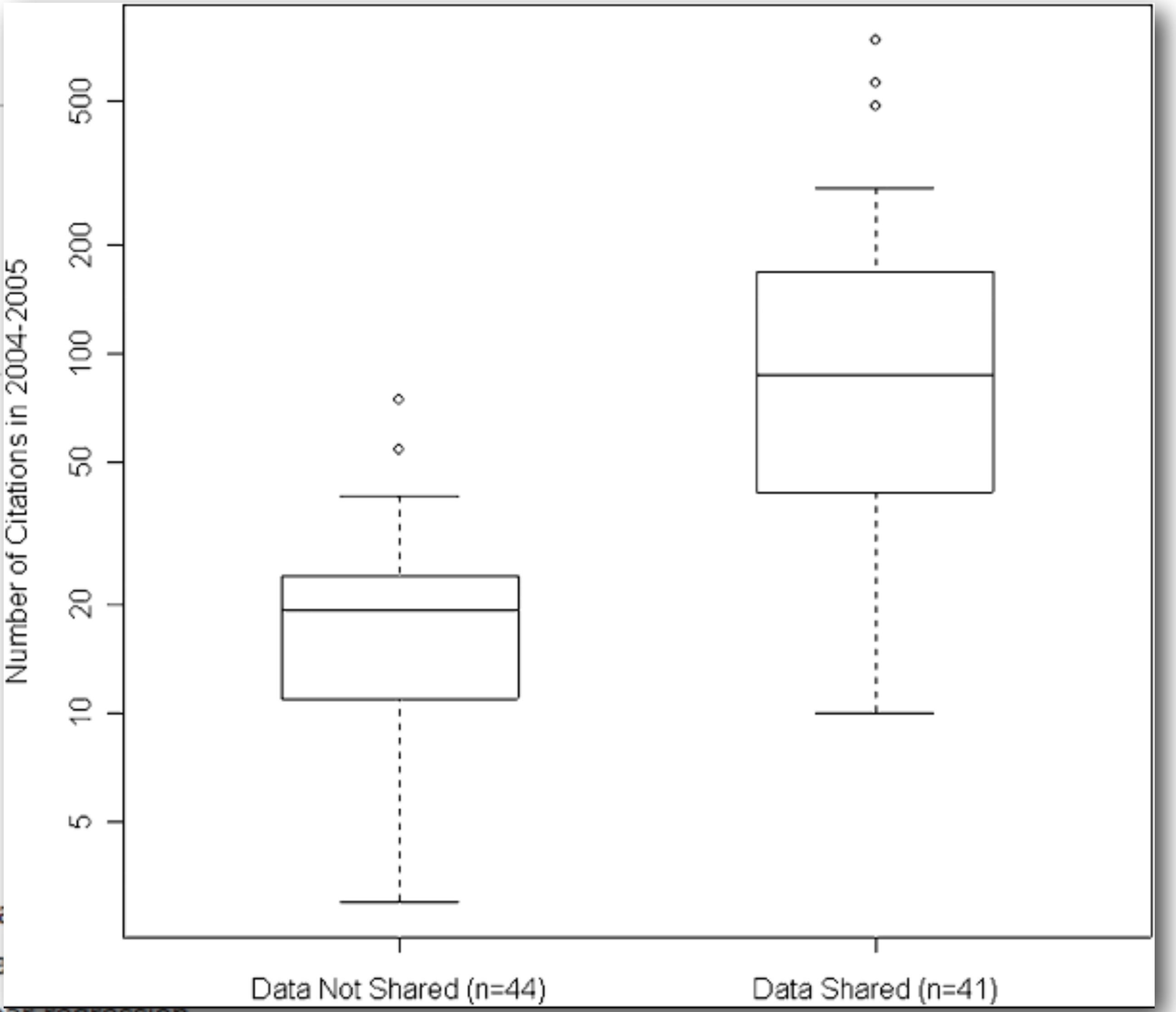
Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available.

Principal Findings

We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggrega
data was significantly (p = 0.006) associated with a 69% increase in cita
impact factor, date of publication, and author country of origin using linear regression.

Significance

This correlation between publicly availab
investigators to share their detailed res



"We found that cancer clinical trials which share their microarray data were cited about 70% more frequently than clinical trials which do not."

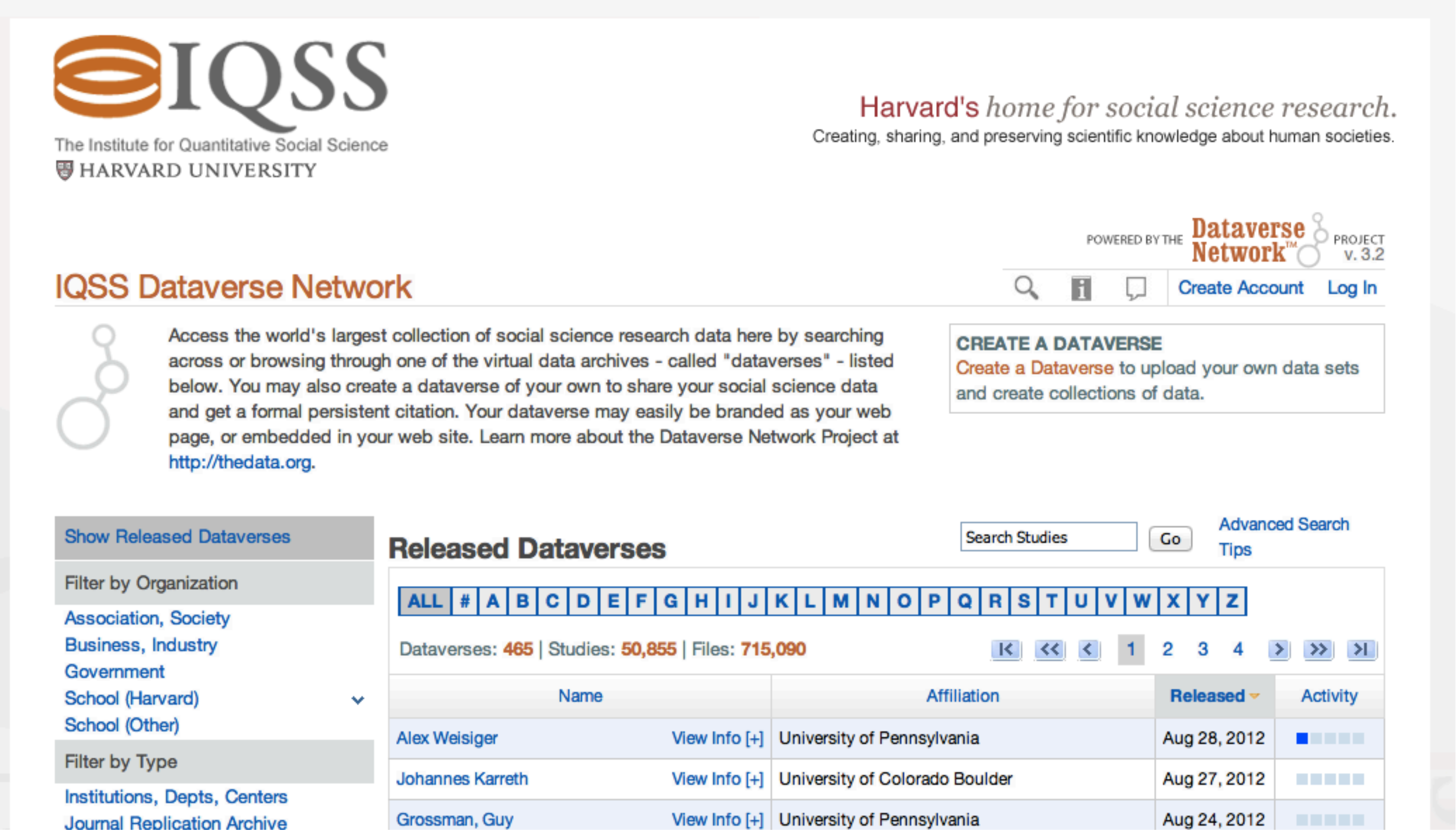
1. Data Sharing and Replication

2. A Solution with the Dataverse Network

open source

The Dataverse Network is a repository for research data that takes care of long term preservation and good archival practices, while researchers keep control of and get recognition for their data.

Dataverse Network for Social Science Data



IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

Harvard's *home for social science research.*

Creating, sharing, and preserving scientific knowledge about human societies.

POWERED BY THE Dataverse Network™ PROJECT v. 3.2

IQSS Dataverse Network

Access the world's largest collection of social science research data here by searching across or browsing through one of the virtual data archives - called "dataverses" - listed below. You may also create a dataverse of your own to share your social science data and get a formal persistent citation. Your dataverse may easily be branded as your web page, or embedded in your web site. Learn more about the Dataverse Network Project at http://thedata.org.

CREATE A DATAVERSE

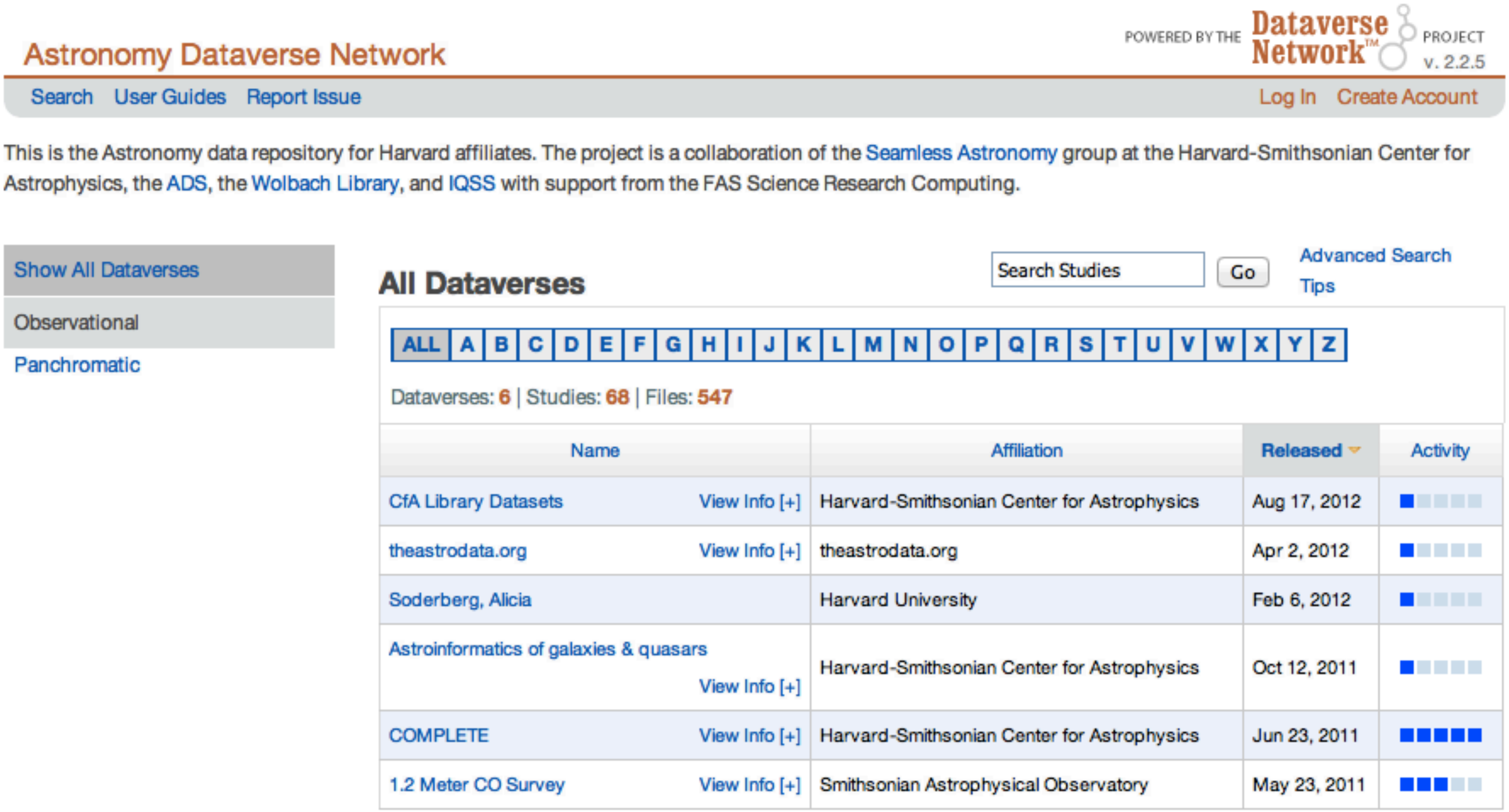
Create a Dataverse to upload your own data sets and create collections of data.

Released Dataverses

Dataverses: 465 | Studies: 50,855 | Files: 715,090

Name		Affiliation	Released	Activity
Alex Weisiger	View Info [+]	University of Pennsylvania	Aug 28, 2012	■□□□□
Johannes Karreth	View Info [+]	University of Colorado Boulder	Aug 27, 2012	□□□□□
Grossman, Guy	View Info [+]	University of Pennsylvania	Aug 24, 2012	□□□□□

Dataverse Network for Astronomy Data



Astronomy Dataverse Network

This is the Astronomy data repository for Harvard affiliates. The project is a collaboration of the Seamless Astronomy group at the Harvard-Smithsonian Center for Astrophysics, the ADS, the Wolbach Library, and IQSS with support from the FAS Science Research Computing.

Observational

Panchromatic

All Dataverses

Dataverses: 6 | Studies: 68 | Files: 547

Name		Affiliation	Released ▾	Activity
CfA Library Datasets	View Info [+]	Harvard-Smithsonian Center for Astrophysics	Aug 17, 2012	■□□□□
theastrodata.org	View Info [+]	theastrodata.org	Apr 2, 2012	■□□□□
Soderberg, Alicia		Harvard University	Feb 6, 2012	■□□□□
Astroinformatics of galaxies & quasars	View Info [+]	Harvard-Smithsonian Center for Astrophysics	Oct 12, 2011	■□□□□
COMPLETE	View Info [+]	Harvard-Smithsonian Center for Astrophysics	Jun 23, 2011	■■■■■
1.2 Meter CO Survey	View Info [+]	Smithsonian Astrophysical Observatory	May 23, 2011	■■■□□

In NeuroInformatics, sharing data for long term access has similar challenges:

✓ **Incentives to share**

✓ Unit of data citation

✓ Sufficient metadata

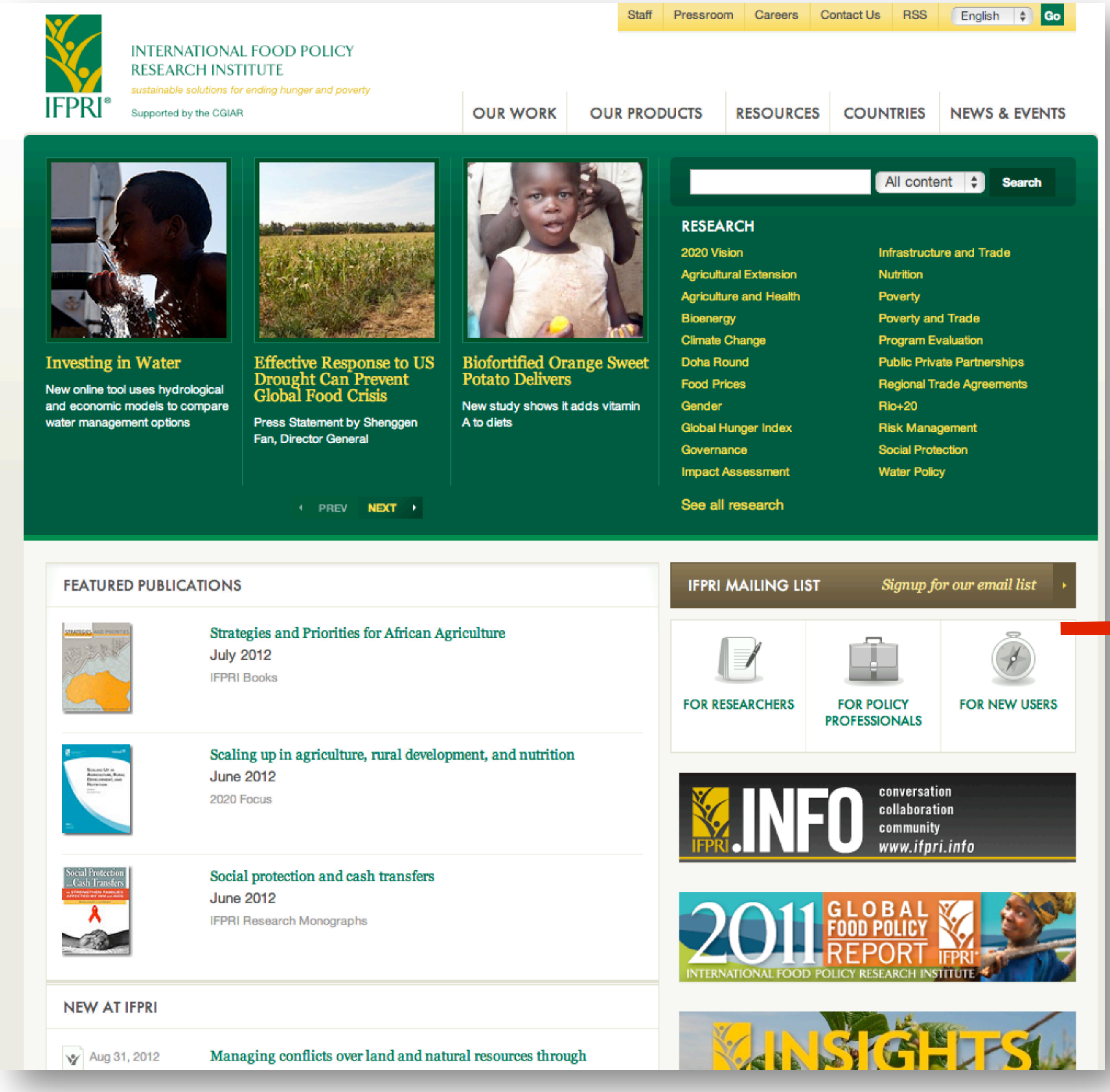
✓ Obsolescence of formats and software

✓ Discoverability

✓ Increasing size of data

✓ ...

Recognition and Credit for Author



Staff | Pressroom | Careers | Contact Us | RSS | English | Go

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

sustainable solutions for ending hunger and poverty

Supported by the CGIAR

IFPRI

OUR WORK | OUR PRODUCTS | RESOURCES | COUNTRIES | NEWS & EVENTS

Investing in Water

New online tool uses hydrological and economic models to compare water management options

Effective Response to US Drought Can Prevent Global Food Crisis

Press Statement by Shenggen Fan, Director General

Biofortified Orange Sweet Potato Delivers

New study shows it adds vitamin A to diets

PREV | NEXT

All content | Search

RESEARCH

2020 Vision
Agricultural Extension
Agriculture and Health
Bioenergy
Climate Change
Doha Round
Food Prices
Gender
Global Hunger Index
Governance
Impact Assessment
Infrastructure and Trade
Nutrition
Poverty
Poverty and Trade
Program Evaluation
Public Private Partnerships
Regional Trade Agreements
Rio+20
Risk Management
Social Protection
Water Policy

See all research

FEATURED PUBLICATIONS

Strategies and Priorities for African Agriculture
July 2012
IFPRI Books

Scaling up in agriculture, rural development, and nutrition
June 2012
2020 Focus

Social protection and cash transfers
June 2012
IFPRI Research Monographs

NEW AT IFPRI

Aug 31, 2012 Managing conflicts over land and natural resources through

IFPRI MAILING LIST *Signup for our email list*

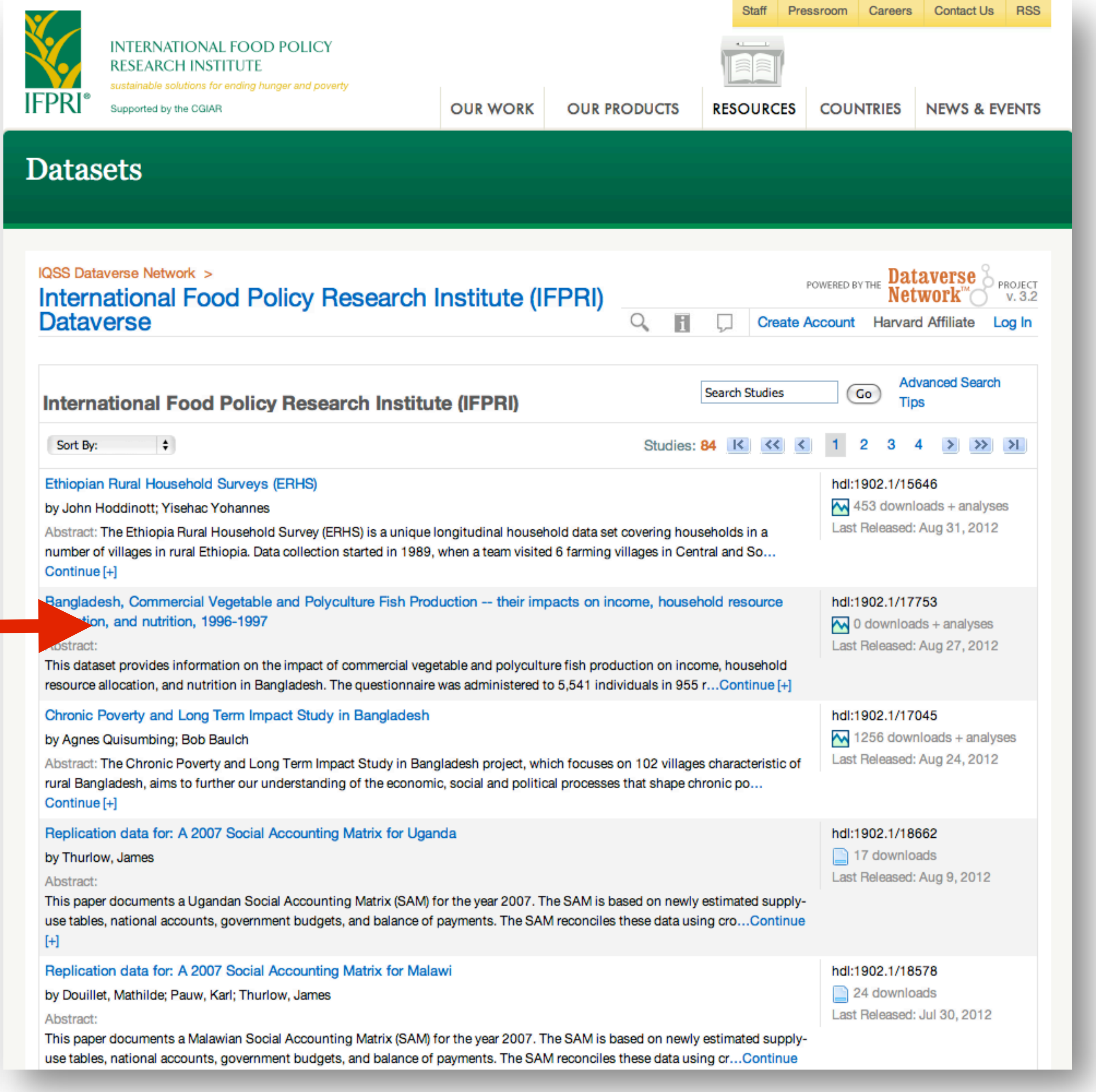
FOR RESEARCHERS | FOR POLICY PROFESSIONALS | FOR NEW USERS

IFPRI .INFO conversation collaboration community www.ifpri.info

2011 GLOBAL FOOD POLICY REPORT IFPRI INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

INSIGHTS

Your own site or project site



Staff | Pressroom | Careers | Contact Us | RSS

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

sustainable solutions for ending hunger and poverty

Supported by the CGIAR

IFPRI

OUR WORK | OUR PRODUCTS | RESOURCES | COUNTRIES | NEWS & EVENTS

Datasets

IQSS Dataverse Network >

International Food Policy Research Institute (IFPRI) Dataverse

POWERED BY THE Dataverse Network™ PROJECT v. 3.2

Create Account | Harvard Affiliate | Log In

International Food Policy Research Institute (IFPRI)

Search Studies | Go | Advanced Search | Tips

Sort By:

Studies: 84 | 1 2 3 4

Ethiopian Rural Household Surveys (ERHS)
by John Hoddinott; Yisehac Yohannes
Abstract: The Ethiopia Rural Household Survey (ERHS) is a unique longitudinal household data set covering households in a number of villages in rural Ethiopia. Data collection started in 1989, when a team visited 6 farming villages in Central and So... Continue [+]

hdl:1902.1/15646
453 downloads + analyses
Last Released: Aug 31, 2012

Bangladesh, Commercial Vegetable and Polyculture Fish Production -- their impacts on income, household resource tion, and nutrition, 1996-1997
bstract:
This dataset provides information on the impact of commercial vegetable and polyculture fish production on income, household resource allocation, and nutrition in Bangladesh. The questionnaire was administered to 5,541 individuals in 955 r...Continue [+]

hdl:1902.1/17753
0 downloads + analyses
Last Released: Aug 27, 2012

Chronic Poverty and Long Term Impact Study in Bangladesh
by Agnes Quisumbing; Bob Baulch
Abstract: The Chronic Poverty and Long Term Impact Study in Bangladesh project, which focuses on 102 villages characteristic of rural Bangladesh, aims to further our understanding of the economic, social and political processes that shape chronic po... Continue [+]

hdl:1902.1/17045
1256 downloads + analyses
Last Released: Aug 24, 2012

Replication data for: A 2007 Social Accounting Matrix for Uganda
by Thurlow, James
Abstract:
This paper documents a Ugandan Social Accounting Matrix (SAM) for the year 2007. The SAM is based on newly estimated supply-use tables, national accounts, government budgets, and balance of payments. The SAM reconciles these data using cro...Continue [+]

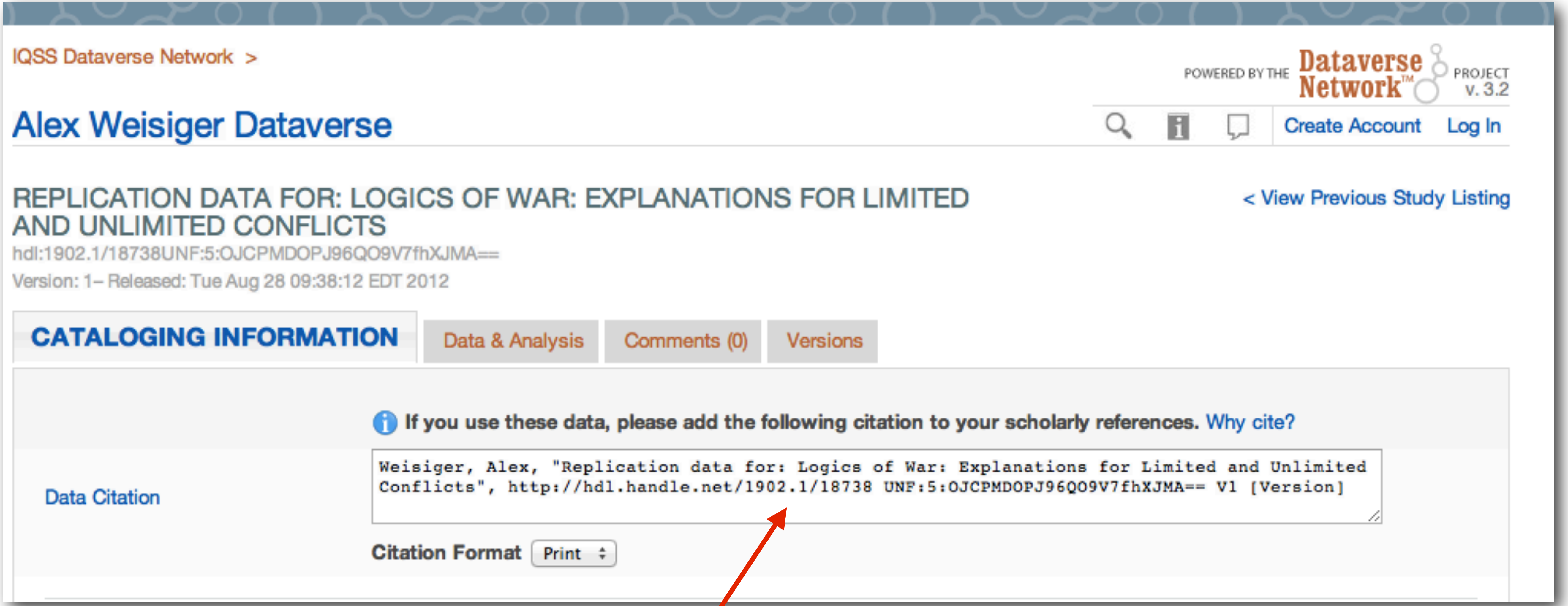
hdl:1902.1/18662
17 downloads
Last Released: Aug 9, 2012

Replication data for: A 2007 Social Accounting Matrix for Malawi
by Douillet, Mathilde; Pauw, Karl; Thurlow, James
Abstract:
This paper documents a Malawian Social Accounting Matrix (SAM) for the year 2007. The SAM is based on newly estimated supply-use tables, national accounts, government budgets, and balance of payments. The SAM reconciles these data using cr...Continue

hdl:1902.1/18578
24 downloads
Last Released: Jul 30, 2012

Your Dataverse

Formal Data Citation



Alex Weisiger Dataverse

REPLICATION DATA FOR: LOGICS OF WAR: EXPLANATIONS FOR LIMITED AND UNLIMITED CONFLICTS

hdl:1902.1/18738UNF:5:OJCPMDOPJ96QO9V7fhXJMA==

Version: 1– Released: Tue Aug 28 09:38:12 EDT 2012

If you use these data, please add the following citation to your scholarly references.

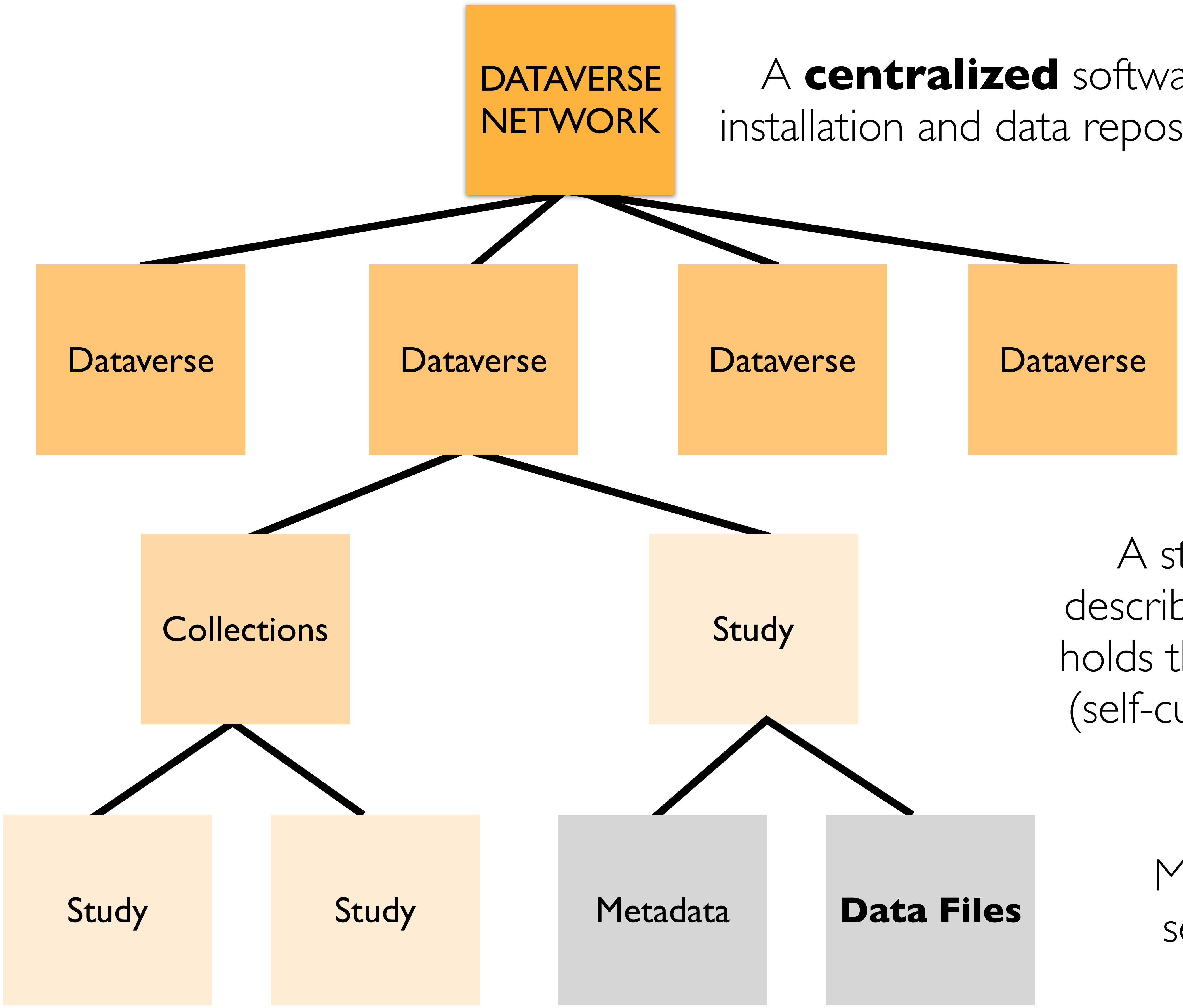
```
Weisiger, Alex, "Replication data for: Logics of War: Explanations for Limited and Unlimited
Conflicts", http://hdl.handle.net/1902.1/18738 UNF:5:OJCPMDOPJ96QO9V7fhXJMA== V1 [Version]
```

Weisiger, Alex, "Replication data for: Logics of War: Explanations for Limited and Unlimited Conflicts", http://hdl.handle.net/1902.1/18738 UNF:5:OJCPMDOPJ96QO9V7fhXJMA== V1 [Version]

Persistent URL (Handle) for permanent reference

Universal Numerical Fingerprint (UNF) for verification

Altman, M., King, G., 2007 "A Proposed Standard for the Scholarly Citation of Quantitative Data"



DATAVERSE NETWORK

A **centralized** software installation and data repository

Dataverse

Dataverse

Dataverse

Dataverse

Individual virtual data archive with its own branding

Collections

Study

A study describes and holds the data (self-curated)

Study

Study

Metadata

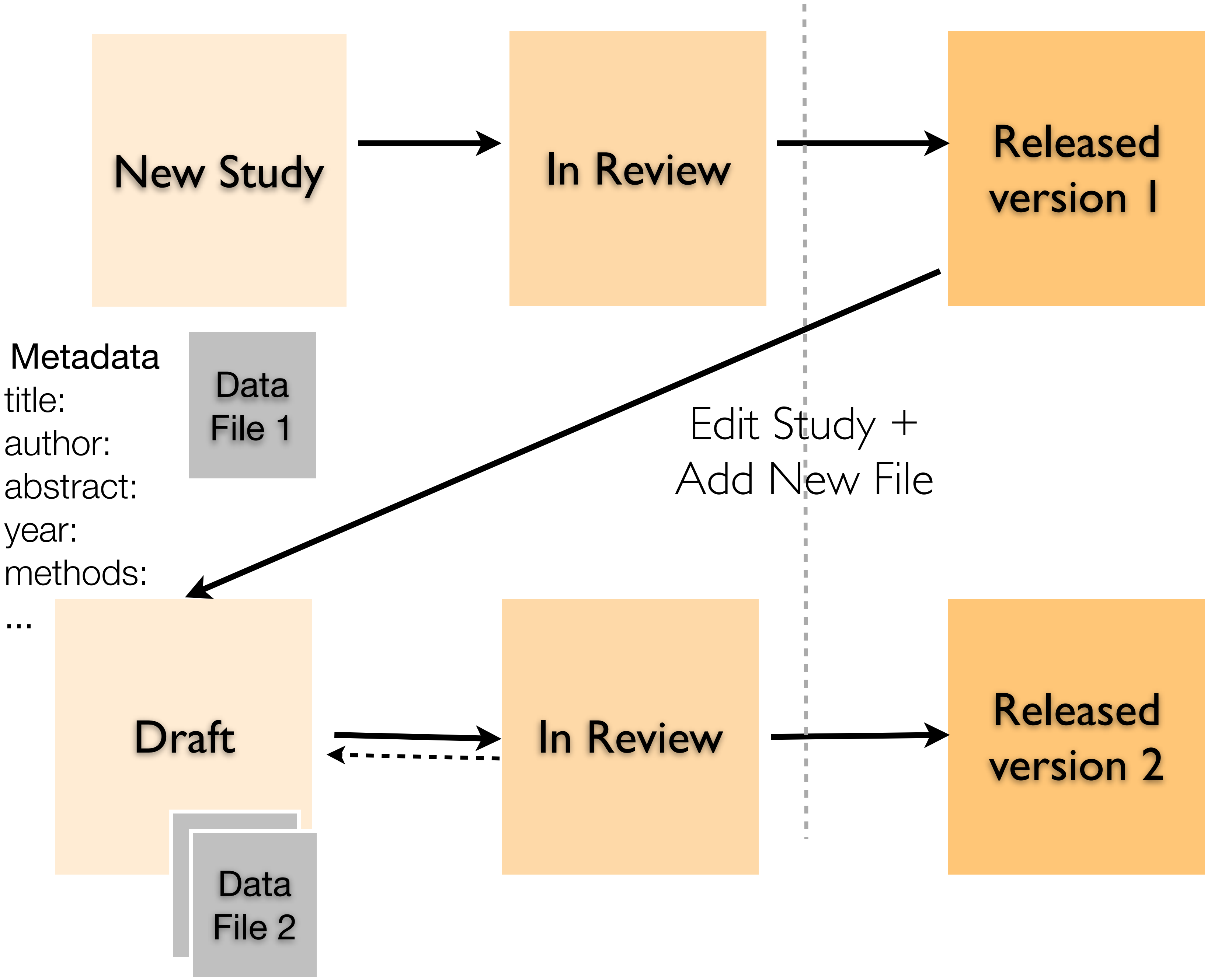
Data Files

Metadata is searchable

Data Versioning and Data Management

Contributors, curators, admins view

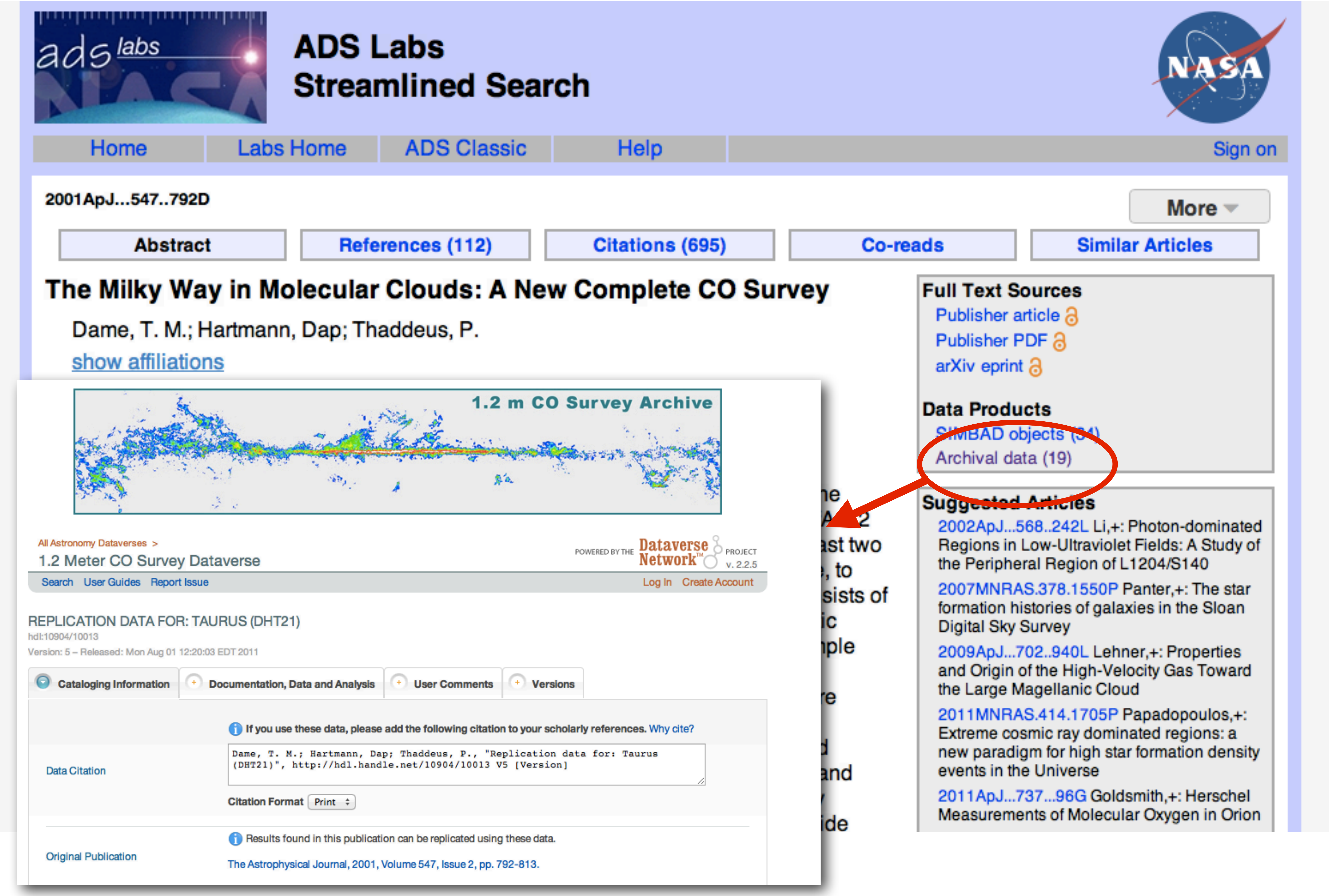
End user view



Christopher Casillas; Peter K. Enns; Patrick C. Wohlfarth, "How Public Opinion Constrains the Supreme Court", http://hdl.handle.net/1902.1/14568 UNF:5:YBYuXzOp6STpakyTEIoScQ== V1 [Version]

Christopher Casillas; Peter K. Enns; Patrick C. Wohlfarth, "How Public Opinion Constrains the Supreme Court", http://hdl.handle.net/1902.1/14568 UNF:5:dI8qi49P0uIB9pLfXA3RCw== V2 [Version]

Connecting Publications to Data



ADS Labs
Streamlined Search

2001ApJ...547..792D

The Milky Way in Molecular Clouds: A New Complete CO Survey

Dame, T. M.; Hartmann, Dap; Thaddeus, P.

show affiliations

he
A 2
ast two
, to
sists of
ic
nple
re
d
and
/
ide

Full Text Sources

- Publisher article
- Publisher PDF
- arXiv eprint

Data Products

- SIMBAD objects (34)
- Archival data (19)

Suggested Articles

- 2002ApJ...568..242L Li,+: Photon-dominated Regions in Low-Ultraviolet Fields: A Study of the Peripheral Region of L1204/S140
- 2007MNRAS.378.1550P Panter,+: The star formation histories of galaxies in the Sloan Digital Sky Survey
- 2009ApJ...702..940L Lehner,+: Properties and Origin of the High-Velocity Gas Toward the Large Magellanic Cloud
- 2011MNRAS.414.1705P Papadopoulos,+: Extreme cosmic ray dominated regions: a new paradigm for high star formation density events in the Universe
- 2011ApJ...737...96G Goldsmith,+: Herschel Measurements of Molecular Oxygen in Orion

1.2 m CO Survey Archive

All Astronomy Dataverses >
1.2 Meter CO Survey Dataverse

REPLICATION DATA FOR: TAURUS (DHT21)
hdl:10904/10013
Version: 5 – Released: Mon Aug 01 12:20:03 EDT 2011

If you use these data, please add the following citation to your scholarly references. Why cite?

Data Citation: Dame, T. M.; Hartmann, Dap; Thaddeus, P., "Replication data for: Taurus (DHT21)", http://hdl.handle.net/10904/10013 V5 [Version]

Citation Format: Print

Results found in this publication can be replicated using these data.

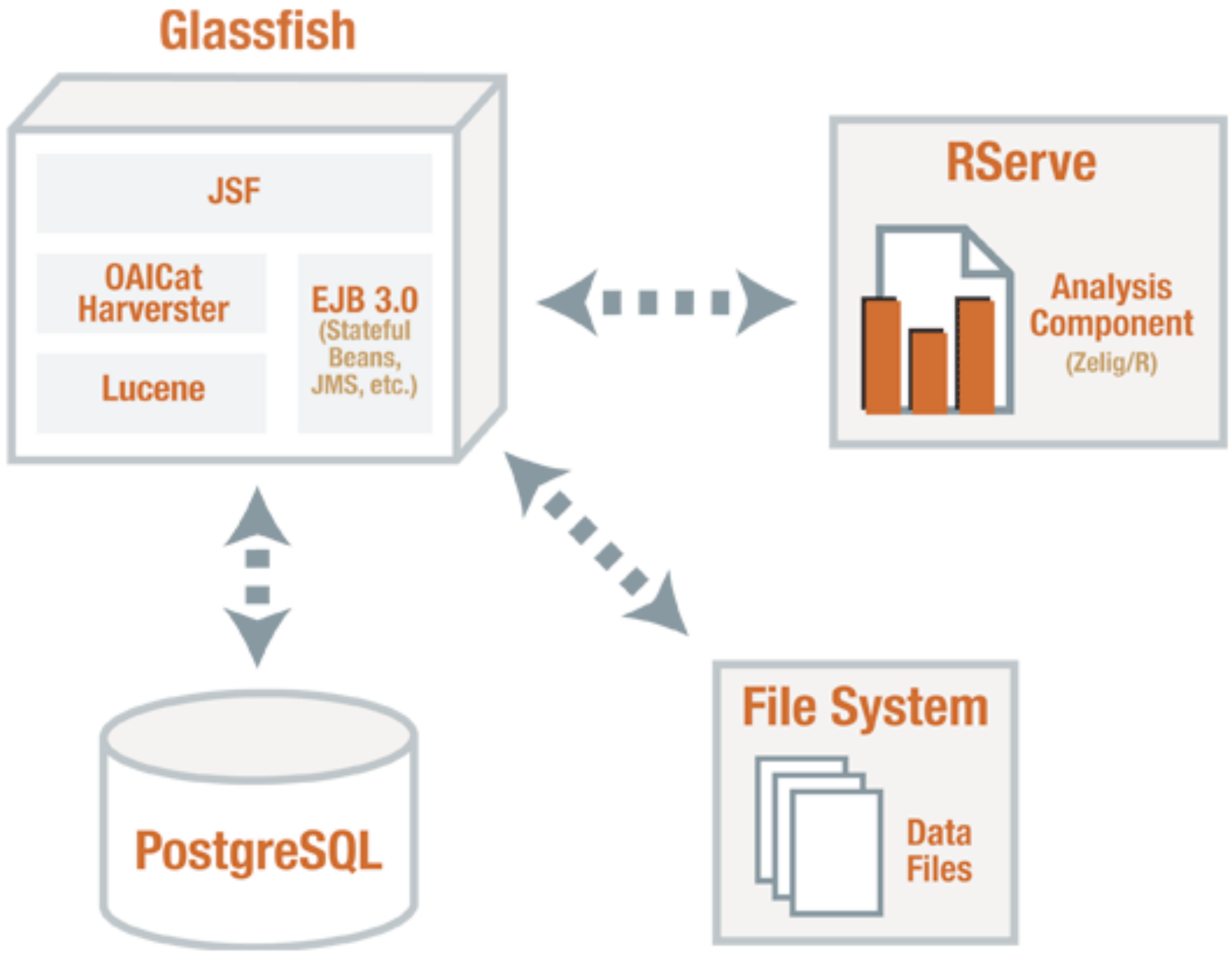
Original Publication: The Astrophysical Journal, 2001, Volume 547, Issue 2, pp. 792-813.

The Centralized Data Repository Provides:

- ✓ Backups and replication of data in different locations (LOCKSS)
- ✓ Re-formatting for preservation (e.g. from SPSS, STATA to archival format)
- ✓ Extraction of Metadata from data sets
- ✓ Metadata standards (DDI, Dublin Core)
- ✓ Inter-operability (OAI-PMH, APIs)

It handles good archival practices for you

Open Source, Java EE 6 Architecture



- JSF, Javascript (jquery): user interface
- EJB: middle-tier, business logic
- OAI-PMH client and server: Harvest Metadata
- Lucene: Indexes metadata
- PostgresSQL: Persistence storage of metadata
- Files system: Store data and complementary files.
- RServe: Analysis component for quantitative data files

concerns and suggestions from the neuroinformatics community?

The Dataverse Network Project: http://thedata.org

mcrosas@iq.harvard.edu
IQSS, Harvard University