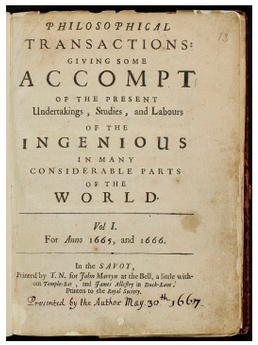
Making Data Accessible

Mercè Crosas
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University

Scholarly output doubles every 20 years from 1750s to 2000

Philosophical Transactions of the Royal Society
Nullius in verba


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol. I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane; Printers to the Royal Society.
Prostant. by the Author May 30th 1667.

Increase in specialization:
every 100-150 authors =
1 new journal for 500-1000 readers

3 journals

10 journals

400 journals

14, 000 journals
(peer-reviewed)

1665 1700 1800 1900 2000

Now:
- 80,000 total journals
- 33,000 peer-reviewed

Source: Mabe, The Growth and Number of Journals, 2003

Data are part of the article in the form of visuals and tables

30% of articles with visuals; 5% have tables

30% with visuals; use of tables increases

50% with visuals, ~7 each article; integrated with tables

90% with figures and tables; occupy 25% of space

1665

1700

First line graphs and bar charts: Playfair (1786)

1800

First pie chart: Playfair (1801)

1900

First scatterplots: Hershel (1833), Galton (1896)

2000

Data visuals evolve from illustrations to scientific arguments

Source: Gross Harmon, Reidy, Communicating Science, 2002

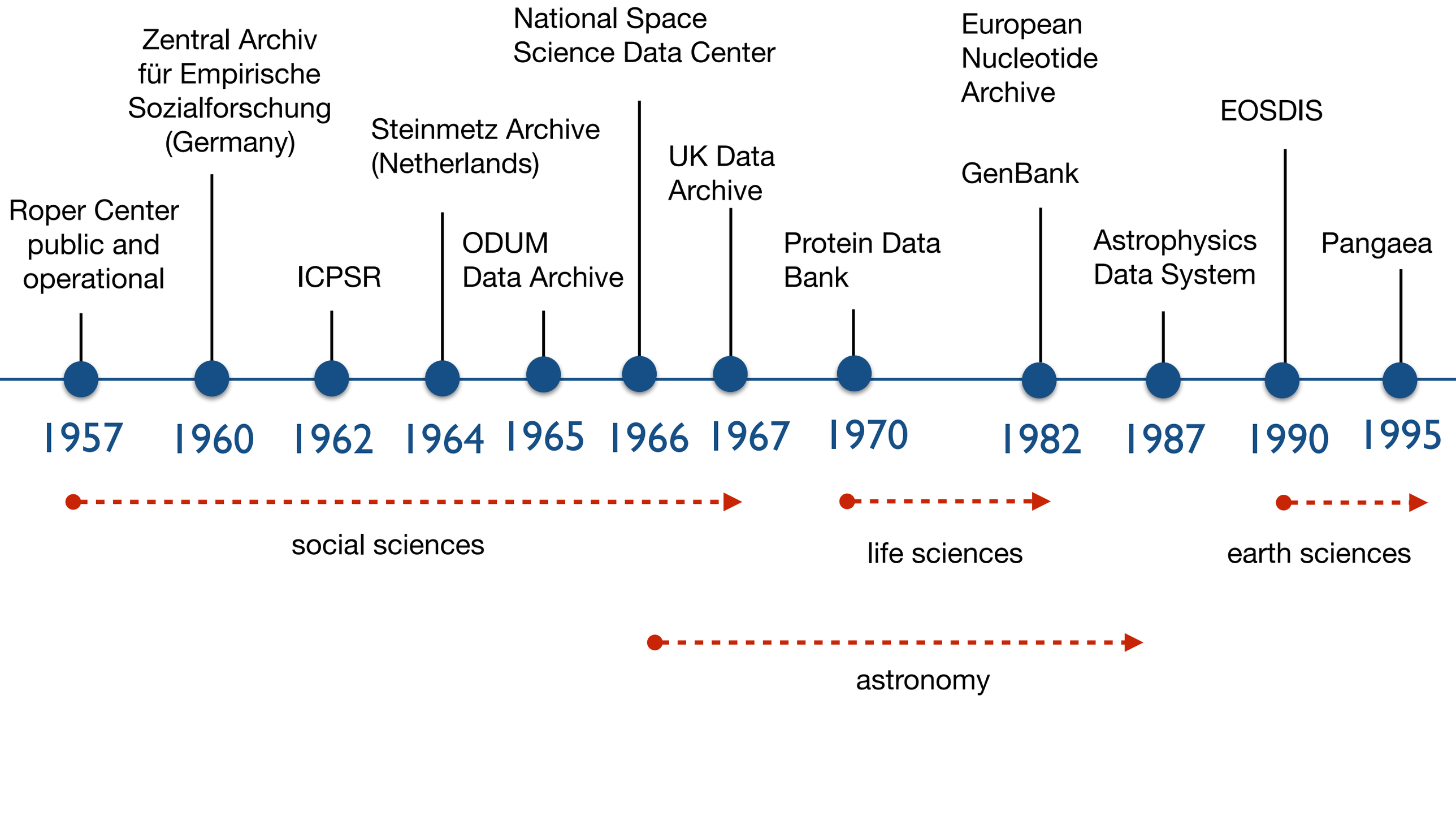
Science communication adapts to the increase in cognitive complexity

- Gross, Herman and Reidy (2002):
  - analyze the style, presentation, and argument of a sample of articles from 17th to 20th century
  - and argue that during that time we developed devices for more efficient communication to compensate the increase of cognitive complexity

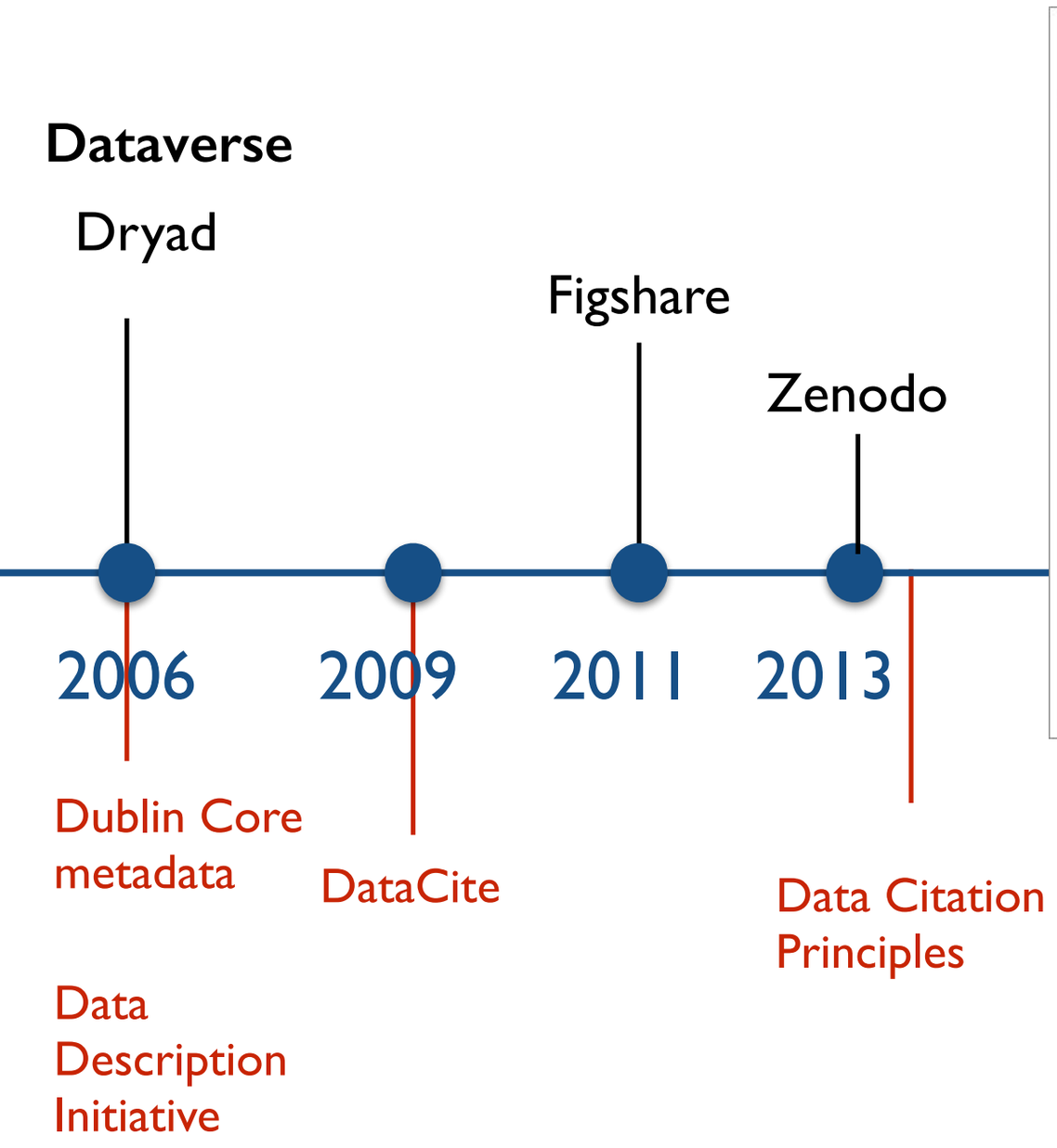
- The increase in quantity and complexity of scholarly output is accompanied with an increase in research data

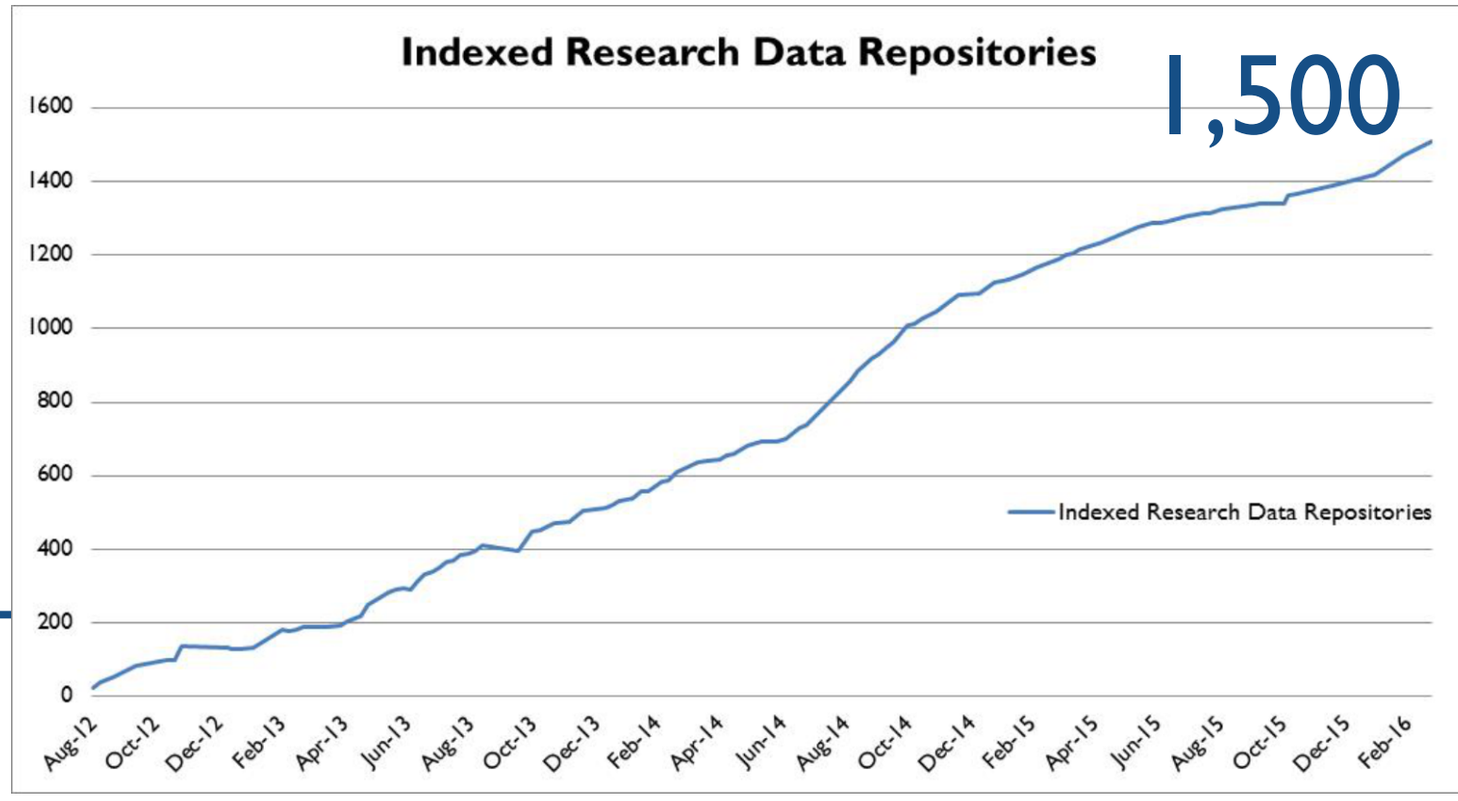
- I argue that in the last decade data publishing is born as a new necessary device for more efficient communication of science

Emergence of (domain-specific) digital data archives



Data publishing in the hands of the researcher





of (all types of) data repositories from 2012 to 2016

source: r3data.org

"Research data publishing is the release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a unique and persistent way."

RDA Data Publishing Workflows Working Group; 10.5281/zenodo.34542

Best Practices for data publishing

- **Data Citation:** to reference and locate with a persistent identifier, and give credit to data authors
- **Metadata:** to discover and reuse
- **Access control rules:** to support publishing workflows and data agreements and licenses, and protect privacy
- **APIs and standards:** to interoperate

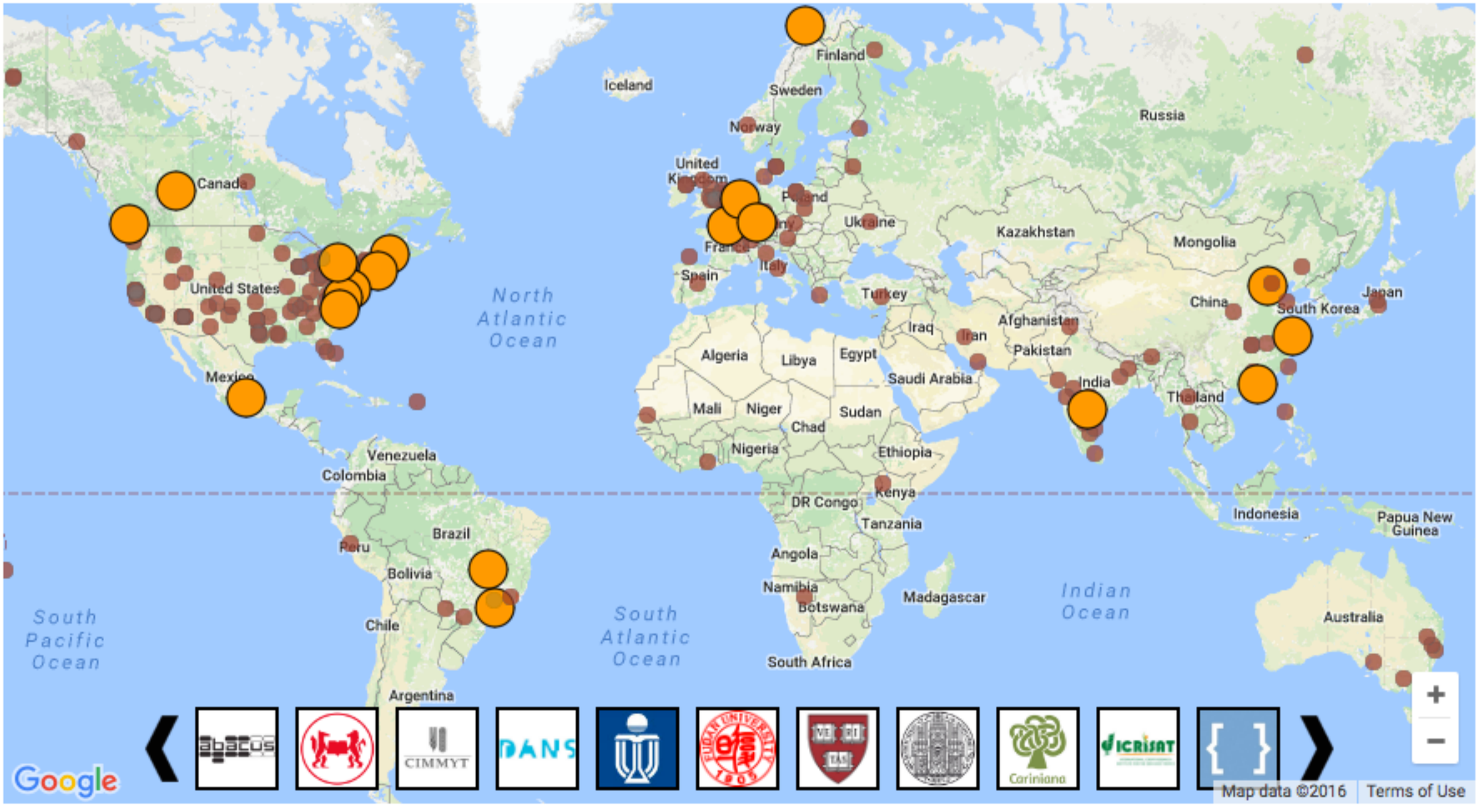
The FAIR Guiding Principles scientific data management and stewardship, 2016; NIH Data Commons Principles

The Role of Dataverse

- Dataverse is an open-source software platform for building data repositories
- 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 Software Today

Installed in 20 sites world wide;
Hosts dataverses from > 500 institutions



The Dataverse Community Today

35

GitHub
contributors

300

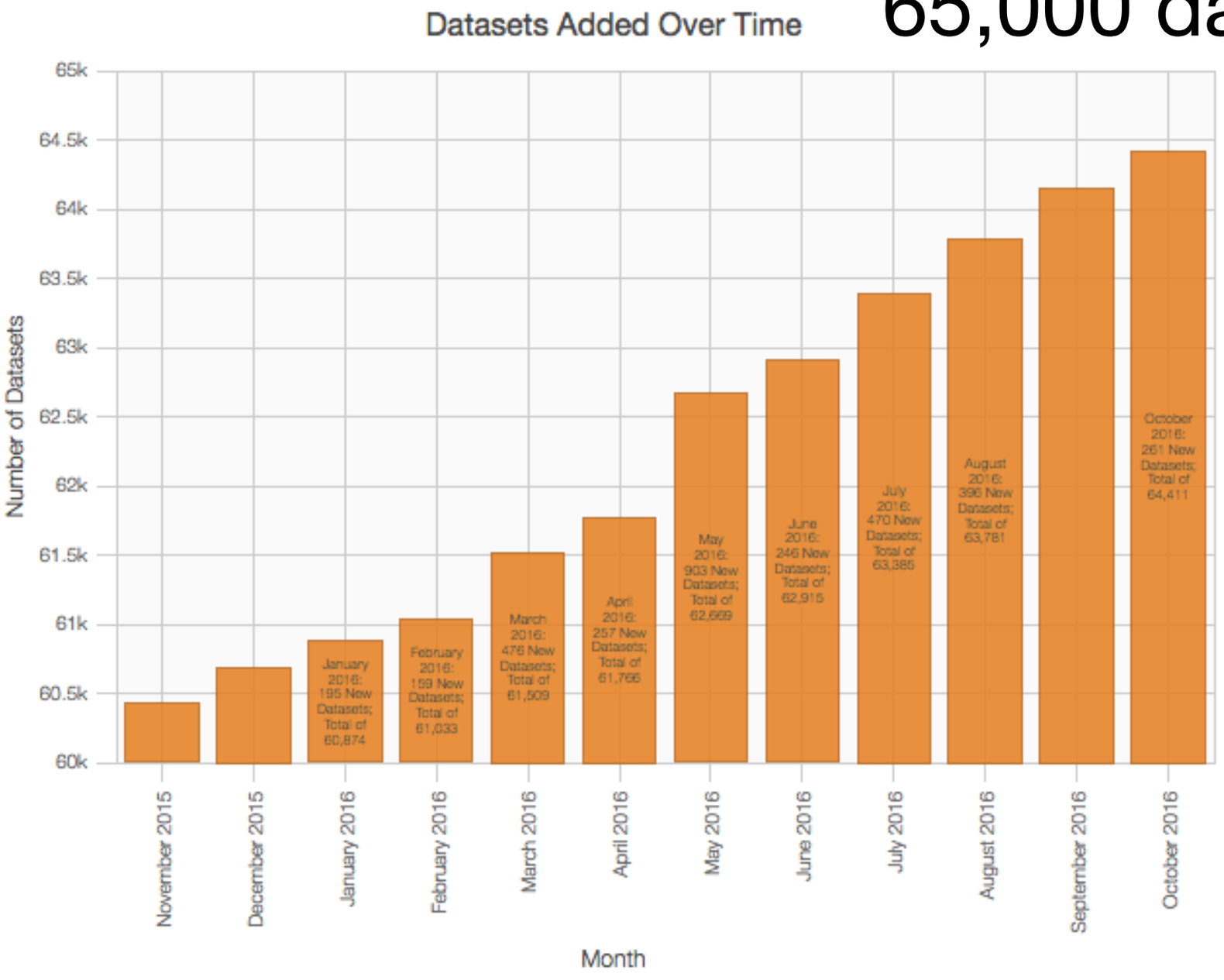
members in the
community list

Semi-monthly
community calls

Annual
Community Meeting,
with near 200
participants

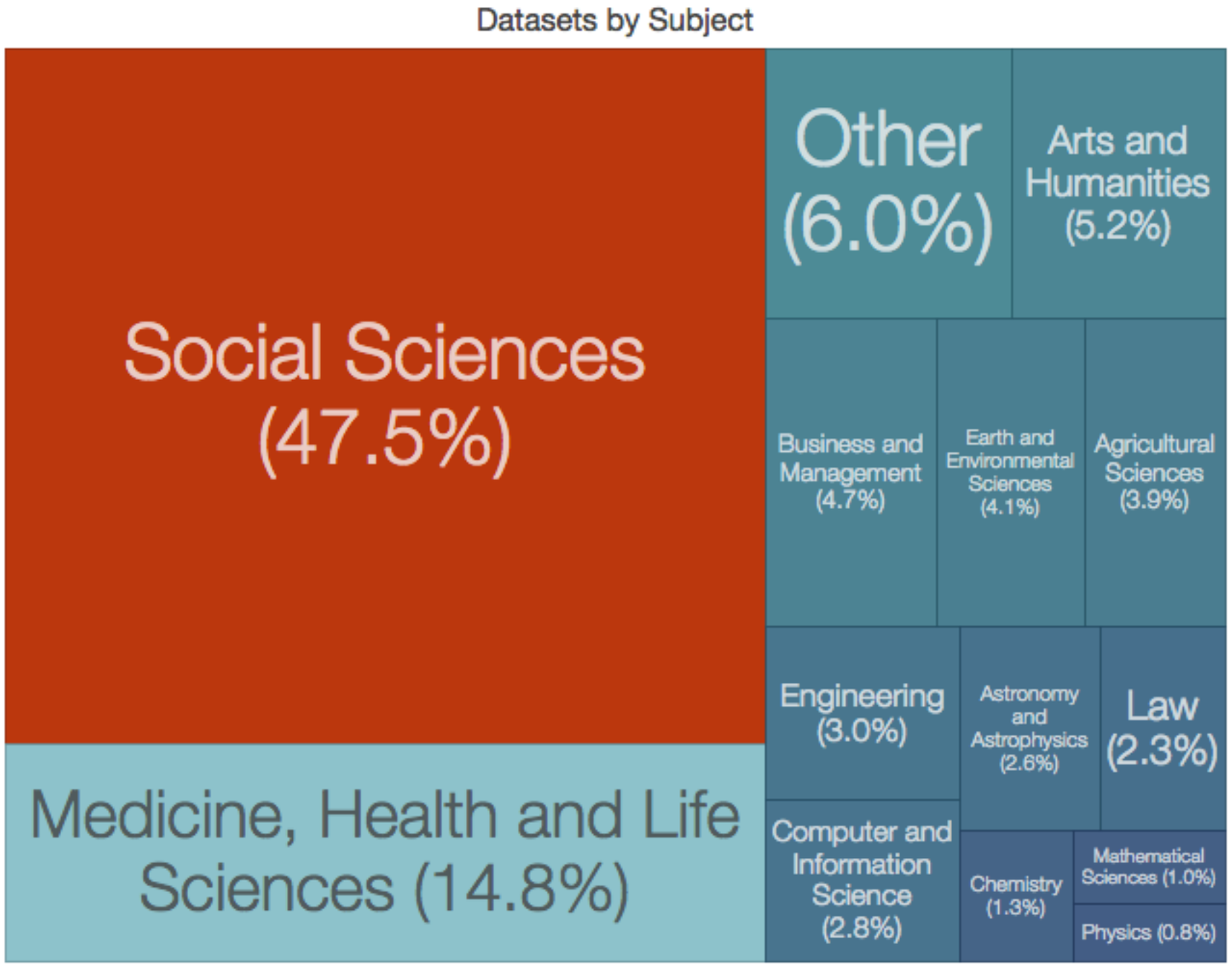
The Harvard Dataverse Today

65,000 datasets

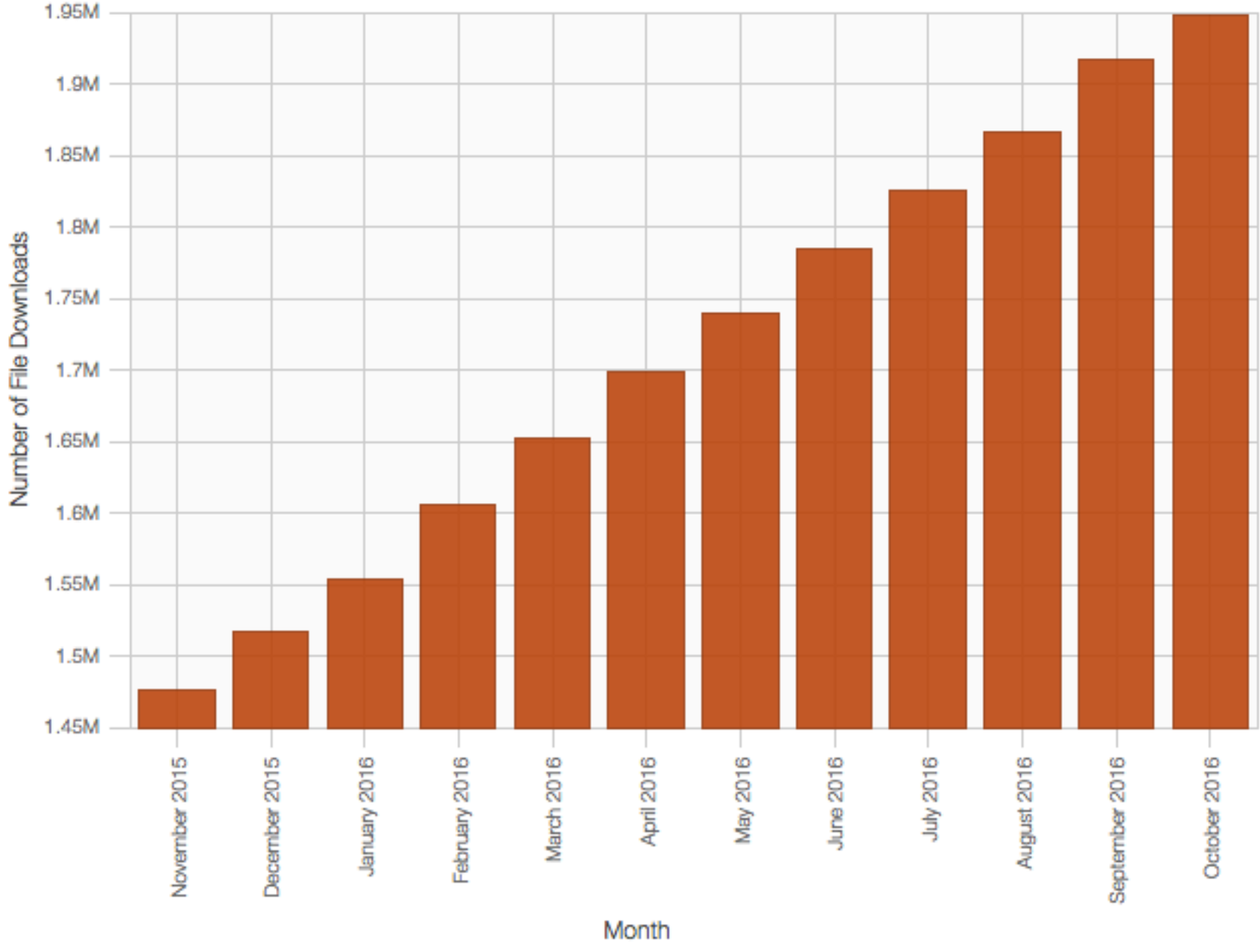


~12 new datasets published per day

The Harvard Dataverse Today



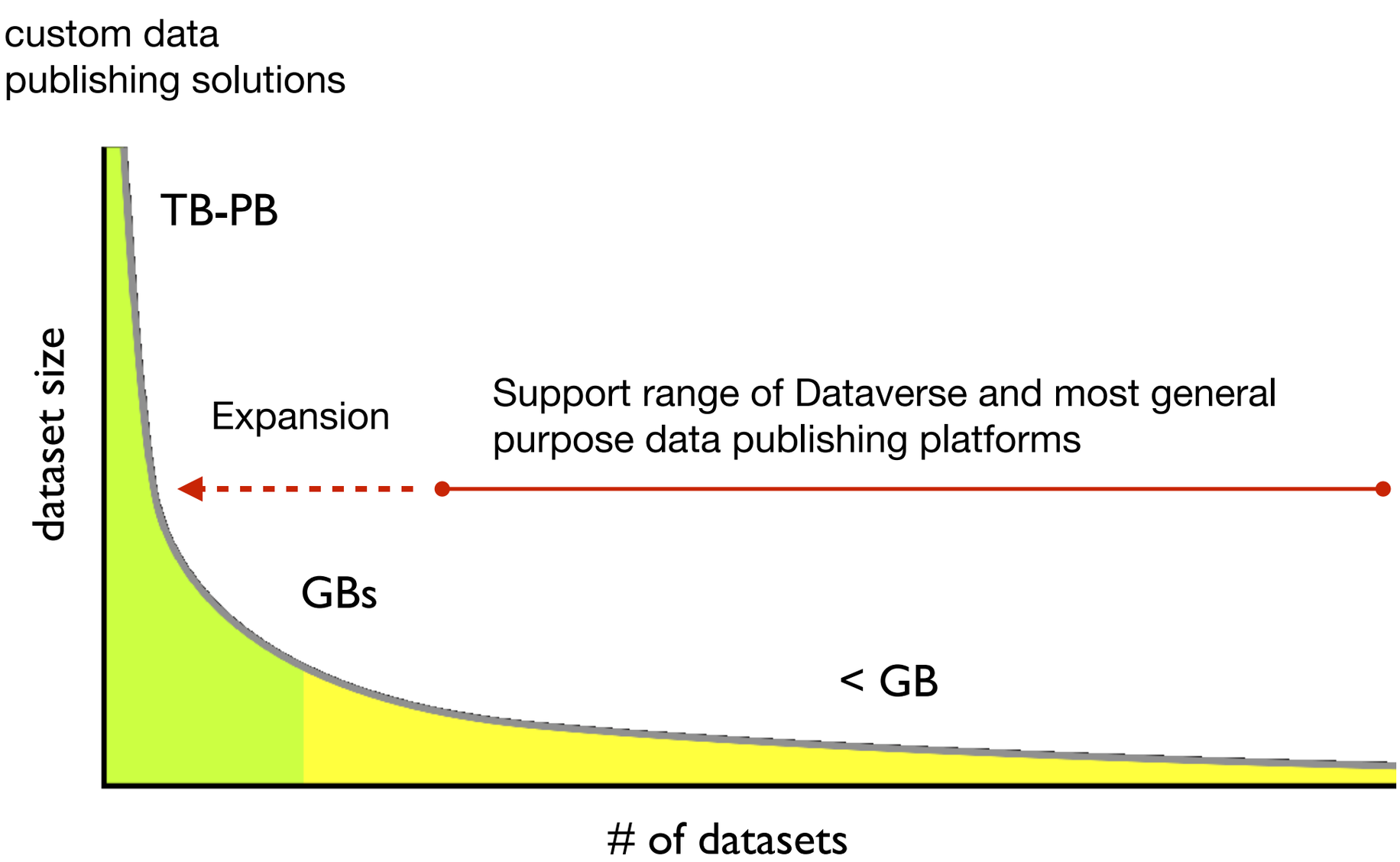
The Harvard Dataverse Today



1.95 Million downloads

~ 1,500 downloads per day

Dataverse Today and Tomorrow



Distribution of Research Data

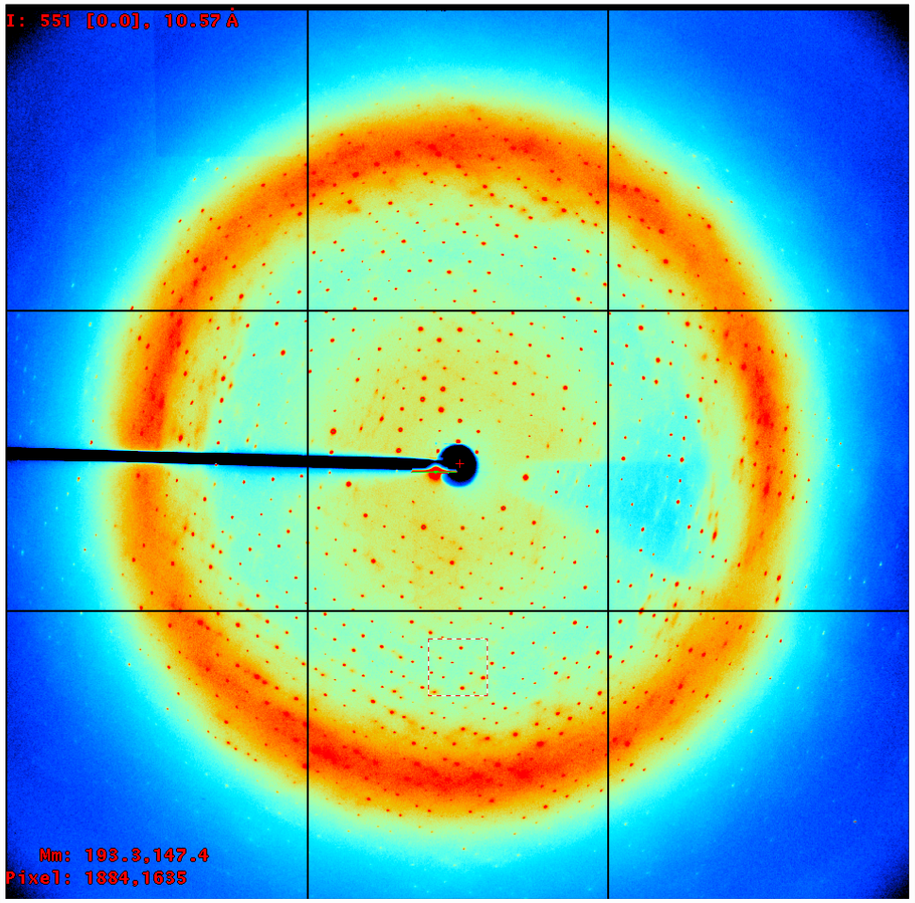
Addressing next challenges

- **Provenance:** Seltzer, Crosas, King
- **Connect journals to data:** King & Crosas
- **Sensitive data:** Harvard Privacy Tools, led by Vadhan
- **Large-scale datasets; integration with remote/cloud storage and computing:**
  - Large datasets in biomedicine, Sliz & Crosas
  - Big Data data in social science, King & Crosas
  - Cloud Dataverse, with Krieger & Saowarattitada (BU)

Making Data Accessible use case: Structural Biology Data Grid

with **Piotr Sliz (Harvard Medical School)**, Pete Meyer, Bill McKinney, Stephanie Socias, Jason Key, Kyle Chard , and IQSS Dataverse team

X-ray diffraction images are the primary data for crystallographic structure determination

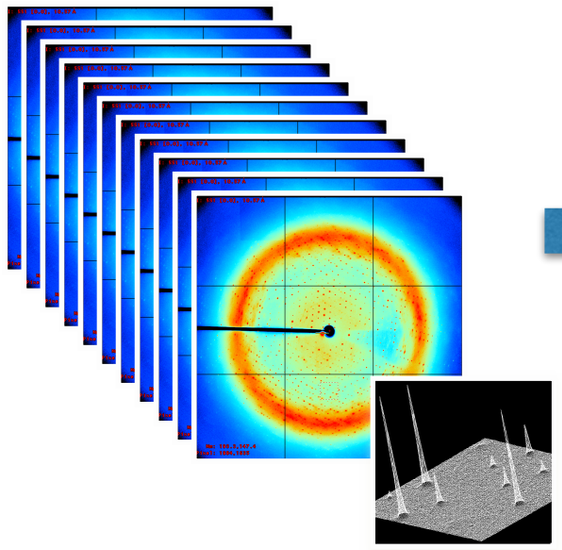


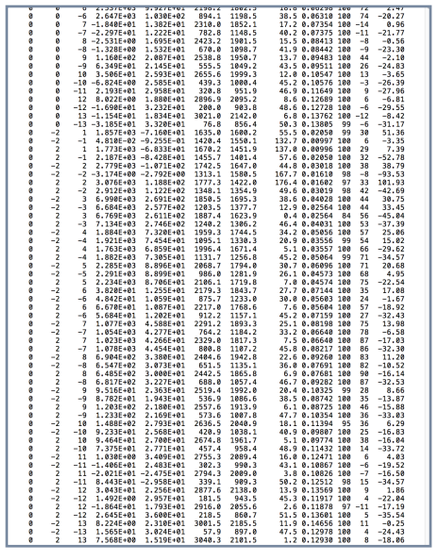
doi:10.15785/SBGRID/19

- X-ray diffraction images collected from frozen protein crystal
- Typically obtained at synchrotron x-ray sources

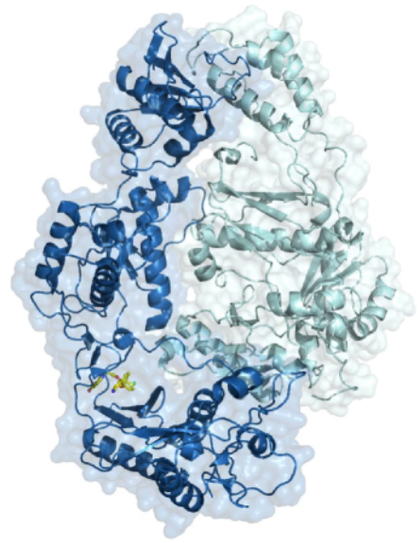
- Primary source for building the protein models published in Protein Data Bank

Why share structural biology primary data?





h	k	l	FP	SIGFP	FC	PHIC	FOM
0	0	-6	2.647E+03	1.030E+02	894.1	1198.5	30.5	0.86318	100	74	-20.27
0	0	7	-1.840E+01	1.382E+01	2310.0	1052.1	17.2	0.87354	100	-14	0.96
0	0	-7	-2.297E+01	1.222E+01	782.8	1148.5	40.2	0.87375	100	-11	-21.77
0	0	8	-2.531E+00	1.695E+01	2423.2	1901.5	15.5	0.88413	100	-8	-0.56
0	0	-8	-1.328E+00	1.532E+01	670.0	1098.7	41.9	0.88442	100	-9	-23.30
0	0	9	1.168E+02	2.807E+01	2538.8	1950.7	13.7	0.89483	100	44	-2.10
0	0	-9	6.349E+01	2.145E+01	555.5	1049.2	43.5	0.89511	100	26	-24.83
0	0	10	3.506E+01	2.593E+01	2655.6	1999.3	12.0	0.10547	100	13	-3.65
0	0	-10	-6.024E+00	2.585E+01	439.3	1000.4	45.2	0.10576	100	-3	-26.39
0	0	-11	2.193E+01	2.958E+01	320.8	951.9	46.9	0.11649	100	9	-27.96
0	0	12	8.022E+00	1.808E+01	2896.9	2095.2	8.6	0.12689	100	6	-6.81
0	0	-12	-1.698E+01	3.232E+01	200.0	903.8	48.6	0.12728	100	-6	-29.55
0	0	13	-1.154E+01	1.834E+01	3021.0	2142.0	6.8	0.13762	100	-12	-8.42
0	0	-13	-3.185E+01	3.328E+01	76.8	856.4	50.3	0.13885	99	-6	-31.17
0	-2	1	1.857E+03	-7.160E+01	1635.0	1600.2	55.5	0.02050	99	30	51.36
0	-2	-1	4.818E-02	-9.255E-01	1420.4	1550.1	132.7	0.00997	100	6	-3.35
0	2	1	1.773E+03	-6.833E+01	1670.2	1451.9	137.0	0.00996	100	29	7.39
0	2	-1	2.187E+03	-8.428E+01	1455.7	1401.4	57.6	0.02050	100	32	-52.78
0	2	2	2.779E+03	-1.071E+02	1742.5	1647.0	44.8	0.03018	100	38	38.79
0	-2	-2	-3.174E+00	-2.792E+00	1313.1	1580.5	167.7	0.01610	98	-8	-93.53
0	2	3	3.076E+03	1.188E+02	1777.3	1422.0	176.4	0.01602	97	33	101.93
0	2	-2	2.912E+03	1.122E+02	1348.1	1354.9	49.6	0.03019	98	42	-42.69
0	-2	3	6.990E+03	2.691E+02	1850.5	1695.3	38.6	0.04028	100	44	30.75
0	-2	-3	6.684E+03	2.577E+02	1203.5	1377.7	12.9	0.02564	84	44	33.45
0	2	3	6.769E+03	2.611E+02	1887.4	1623.9	0.4	0.02564	84	56	-45.04
0	2	-3	7.134E+03	2.746E+02	1248.2	1386.2	46.4	0.04031	100	53	-37.39
0	-2	4	1.884E+03	7.320E+01	1959.3	1744.5	34.2	0.05056	100	57	25.06
0	-2	-4	1.921E+03	7.454E+01	1095.1	1330.3	20.9	0.03556	99	54	15.02
0	2	4	1.763E+03	6.859E+01	1996.4	1671.4	5.1	0.03557	100	66	-29.62
0	2	-4	1.882E+03	7.305E+01	1131.7	1256.8	45.2	0.05064	99	71	-34.57
0	-2	5	2.285E+03	8.896E+01	2068.7	1794.0	30.7	0.06096	100	71	20.68
0	-2	-5	2.291E+03	8.899E+01	986.0	1281.9	26.1	0.04573	100	68	4.95
0	2	5	2.224E+03	8.706E+01	2106.1	1719.1	7.0	0.04574	100	75	-22.54
0	2	-5	3.020E+01	1.255E+01	2179.3	1843.7	27.7	0.07144	100	35	17.08
0	-2	-6	4.842E+01	1.859E+01	875.7	1233.0	30.0	0.05603	100	24	-1.67
0	2	6	6.678E+01	1.807E+01	2217.0	1768.6	7.6	0.05604	100	57	-18.92
0	2	-6	5.684E+01	1.202E+01	912.2	1157.1	45.2	0.07159	100	27	-32.43
0	-2	7	1.077E+03	4.588E+01	2201.2	1893.3	25.1	0.08198	100	75	13.98
0	-2	-7	1.054E+03	4.277E+01	764.2	1184.2	33.2	0.06640	100	78	-6.58
0	2	7	1.023E+03	4.266E+01	2329.0	1817.3	7.5	0.06640	100	87	-17.03
0	2	-7	1.078E+03	4.454E+01	800.8	1187.2	45.8	0.08217	100	86	-32.30
0	-2	8	6.904E+02	3.308E+01	2404.6	1942.8	22.6	0.09260	100	83	11.20
0	-2	-8	6.547E+02	3.073E+01	651.5	1135.1	36.8	0.07691	100	82	-10.52
0	2	8	6.485E+02	3.000E+01	2442.5	1865.8	6.9	0.07681	100	90	-16.14
0	2	-8	6.817E+02	3.227E+01	680.0	1057.4	46.7	0.09282	100	87	-32.53
0	2	9	9.516E+01	2.363E+01	2519.4	1992.0	20.4	0.10325	99	28	8.66
0	-2	-9	8.782E+01	1.943E+01	536.9	1086.6	38.5	0.08742	100	35	-13.87
0	2	9	1.203E+02	2.108E+01	2557.6	1913.9	6.1	0.08725	100	46	-15.88
0	2	-9	1.233E+02	2.169E+01	573.6	1087.8	47.7	0.10354	100	36	-33.03
0	-2	10	1.488E+02	2.793E+01	2636.5	2040.9	18.1	0.11394	95	36	6.29
0	-2	-10	9.233E+01	2.568E+01	420.9	1030.1	40.9	0.09807	100	25	-16.83
0	2	10	9.464E+01	2.700E+01	2674.8	1961.7	5.1	0.09774	100	38	-16.04
0	2	-10	7.375E+01	2.771E+01	457.4	950.4	48.9	0.11432	100	14	-33.72
0	-2	11	1.038E+00	3.409E+01	2755.3	2089.4	16.0	0.12471	100	6	4.03
0	-2	-11	-1.406E+01	2.483E+01	302.3	990.3	43.1	0.10867	100	-6	-19.52
0	2	11	-2.021E+01	-2.475E+01	2794.3	2089.0	3.8	0.10826	100	-7	-16.50
0	2	-11	8.443E+01	-2.958E+01	339.1	909.3	50.2	0.12512	98	15	-34.57
0	-2	12	3.043E+01	2.256E+01	2877.6	2130.0	13.9	0.13569	100	9	1.86
0	-2	-12	1.492E+00	2.957E+01	181.5	943.5	45.3	0.11917	100	4	-22.04
0	2	12	-1.864E+01	1.793E+01	2916.0	2055.6	2.6	0.11878	97	-11	-17.19
0	2	-12	2.645E+01	3.600E+01	218.5	860.7	51.5	0.13601	100	5	-35.54
0	-2	13	8.224E+00	2.310E+01	3001.5	2185.5	11.9	0.14656	100	11	-0.25
0	-2	-13	1.565E+01	3.824E+01	57.9	897.0	47.5	0.12978	100	4	-24.43
0	2	13	7.568E+00	1.519E+01	3040.3	2101.5	1.2	0.12930	100	8	-18.06



1 dataset is 180 - 360 images
3.5 - 7 GB
Primary Data

1 file of ‘structure factors:
0.0001 -0.005 GB
Intermediate Data

Atomic Coordinates
0.0001 -0.001 GB
Model

WORLDWIDE PDB PROTEIN DATA BANK

- Replicate published model results
- Reprocess primary data with new, improved analysis

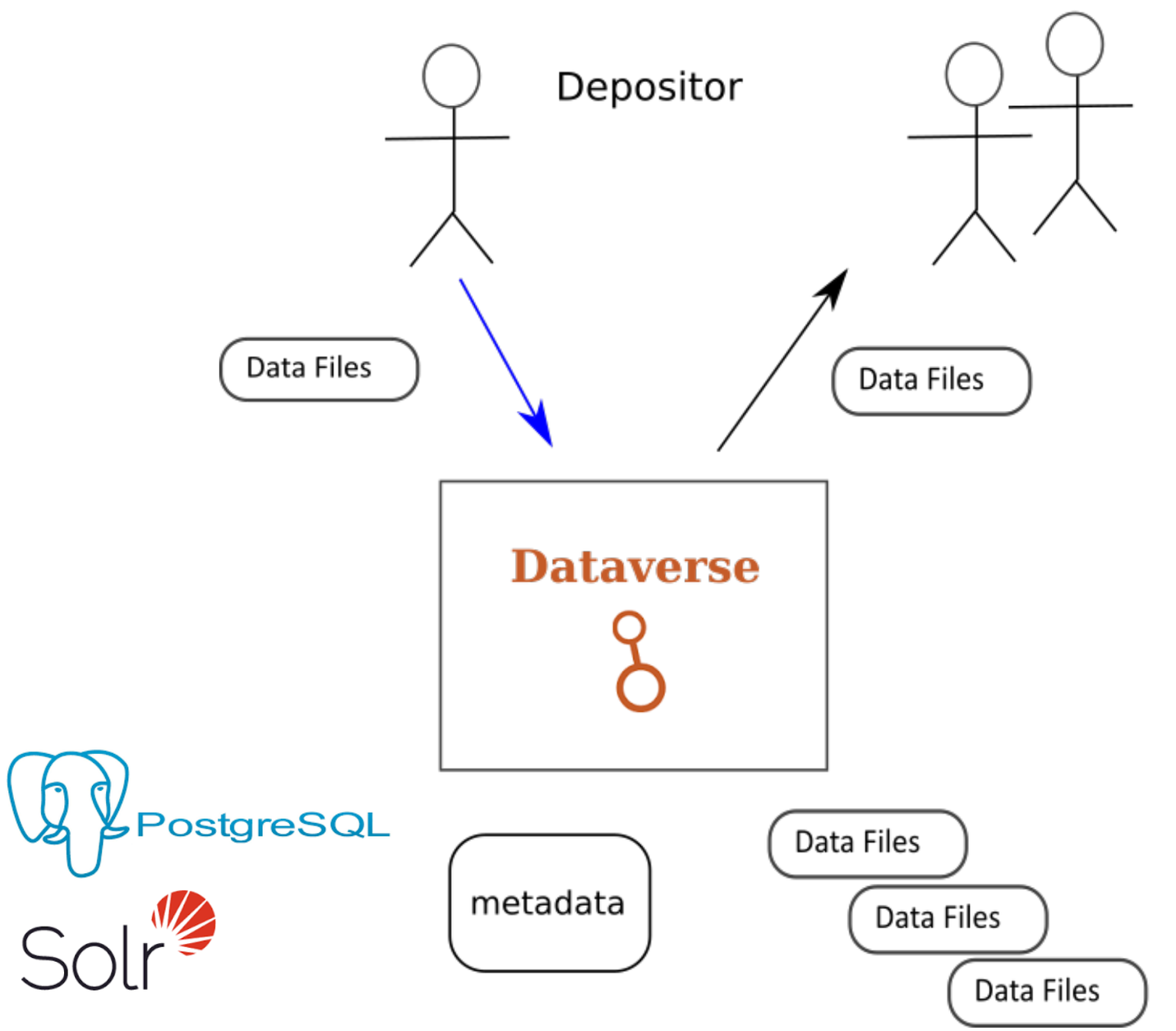
Why Structural Biology Data Grid + Dataverse?

- Support data publication, data access, preservation, and live analysis for primary structural biology data

- Apply best practices in data sharing: data citation, metadata, standards and protocols

- Use an open-source platform with a growing community as a sustainable infrastructure for the repository

Dataverse dataflow now

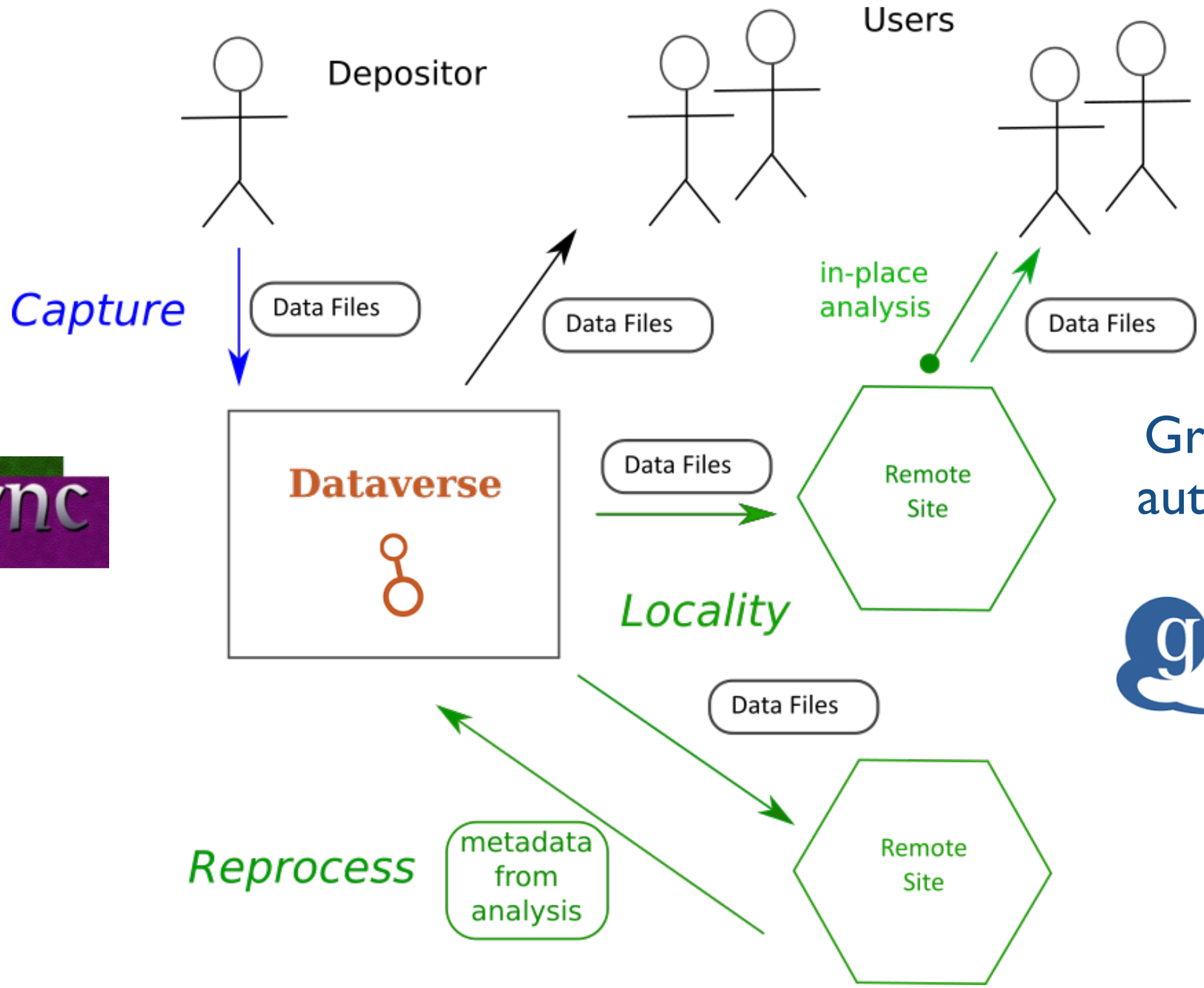


Deposit and get data and metadata through UI, API, and OAI-PMH

http://

https://

Dataverse expansion



rsync

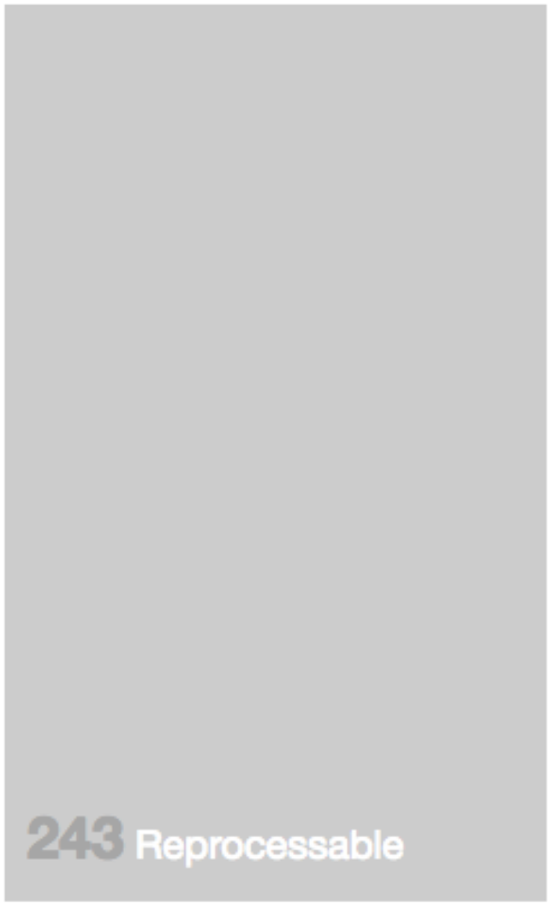
GridFTP with authentication through

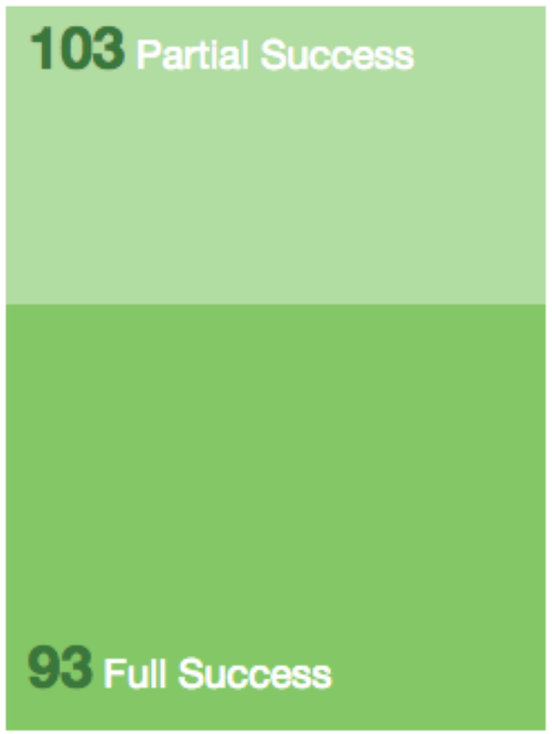
globus

Replication Success Rate

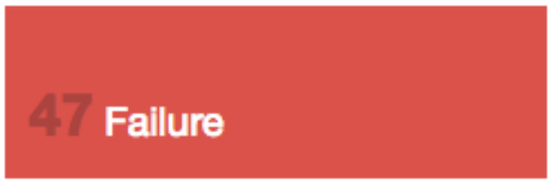
255 Datasets Published

Overall results for datasets with an available reprocessing pipeline shown below. Detailed statistics are reported on dataset landing pages.





85 % success



0 Pending

Reprocessing Pipeline

X-ray reprocessing summary

Dataset	xia2 -2d	xia2 -3d	xia2 -3dii	xia2 -dials
1	✅	✅	✅	✅
2	✅	✅	✅	✅
3	❌	✅	✅	✅
4	✅	✅	✅	✅
5	❌	❌	❌	❌
6	❌	❌	❌	❌
7	❌	✅	❌	✅
9	❌	⛔	⛔	❌
10	❌	❌	❌	✅
11	✅	✅	✅	✅
12	✅	✅	✅	✅
13	❌	❌	❌	❌
14	✅	✅	✅	✅
15	❌	✅	✅	✅

Sliz Laboratory Dataverse (Harvard Medical School)

X-Ray Diffraction data from Lin28/let-7d microRNA complex, source of 3TRZ structure

Sliz, Piotr, 2016, "X-Ray Diffraction data from Lin28/let-7d microRNA complex, source of 3TRZ structure", doi:10.15785/SBGRID/3, Root Dataverse, V1

Cite Data

Learn about Data Citation Standards.

Related Publication	Nam Y, Chen C, Gregory RI, Chou JJ, Sliz P. Molecular Basis for Interaction of let-7 MicroRNAs with Lin28. Cell 2011; 147:1080–1091. doi: 10.1016/j.cell.2011.10.020
3TRZ Coordinates	PDB, MMDB
Biological Sample	Lin28/let-7d microRNA complex
Dataset Type	X-Ray Diffraction
Subject Composition	RNA
Data Creation Date	2010-09-07

Reprocessing Data

>_ xia2 -dials /programs/datagrid/3 verbose=True multiprocessing.nproc=1 -ispyb_xml_out apc.xml

>_ xia2 -3dii /programs/datagrid/3 verbose=True multiprocessing.nproc=1 -ispyb_xml_out apc.xml

>_ xia2 -2d /programs/datagrid/3 verbose=True multiprocessing.nproc=1 -ispyb_xml_out apc.xml

Version: 0.4.0.0

Reprocessing failed.

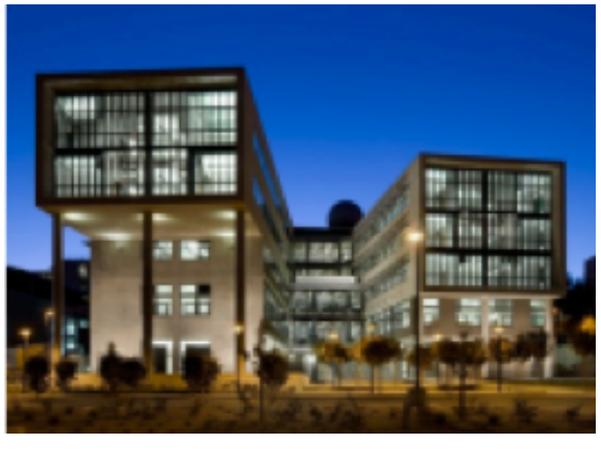
>_ xia2 -3d /programs/datagrid/3 verbose=True multiprocessing.nproc=1 -ispyb_xml_out apc.xml

Data Access Alliance:

Local Access through a growing list of satellites

Harvard Medical School Orchestra Cluster

Petrel Storage
Argonne Labs



San Diego Supercomputer Center

UPPSALA UNIVERSITET

Institut Pasteur de Montevideo

SIBS

SHANGHAI INSTITUTES FOR BIOLOGICAL SCIENCES

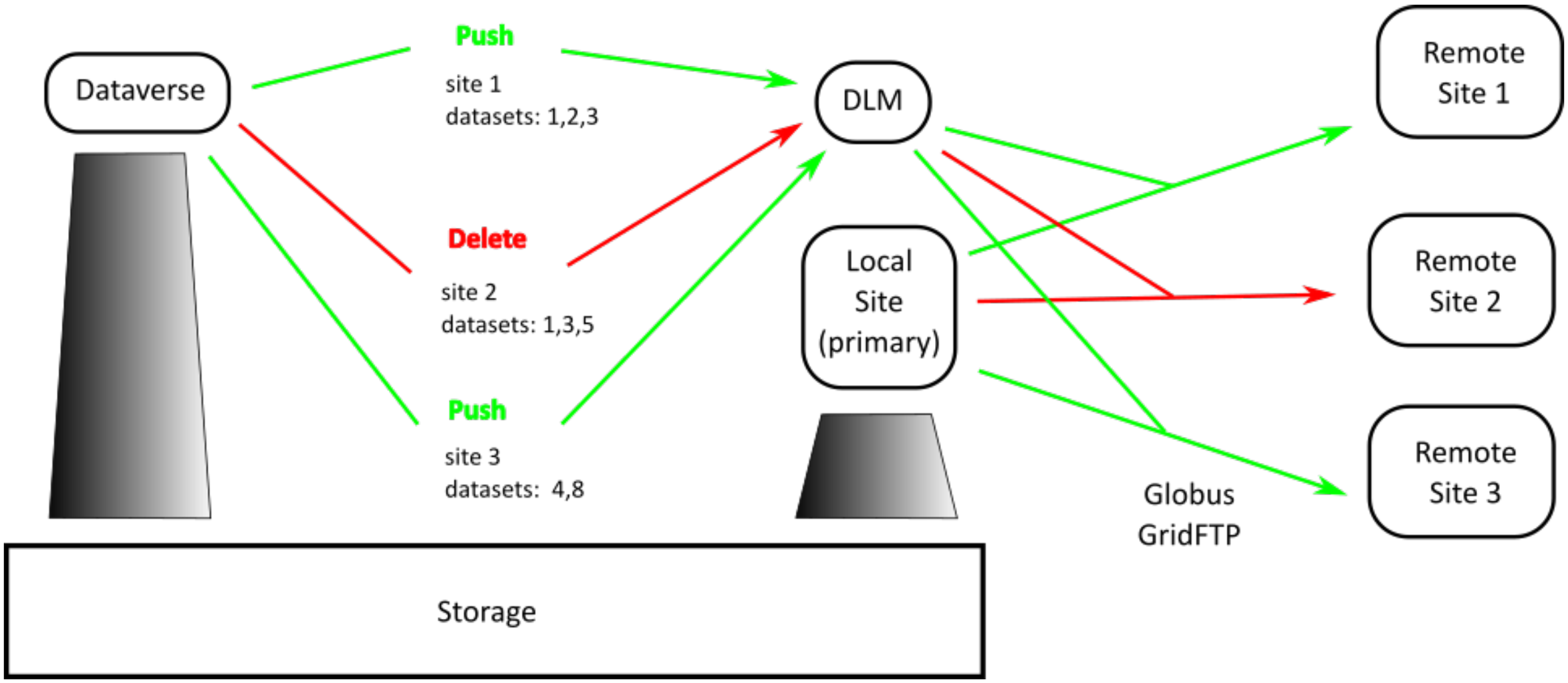
Meyer et al., Nature Communications, 2016

Data Access Alliance Flow

OAI set per remote site

Data transfer through GridFTP

globus

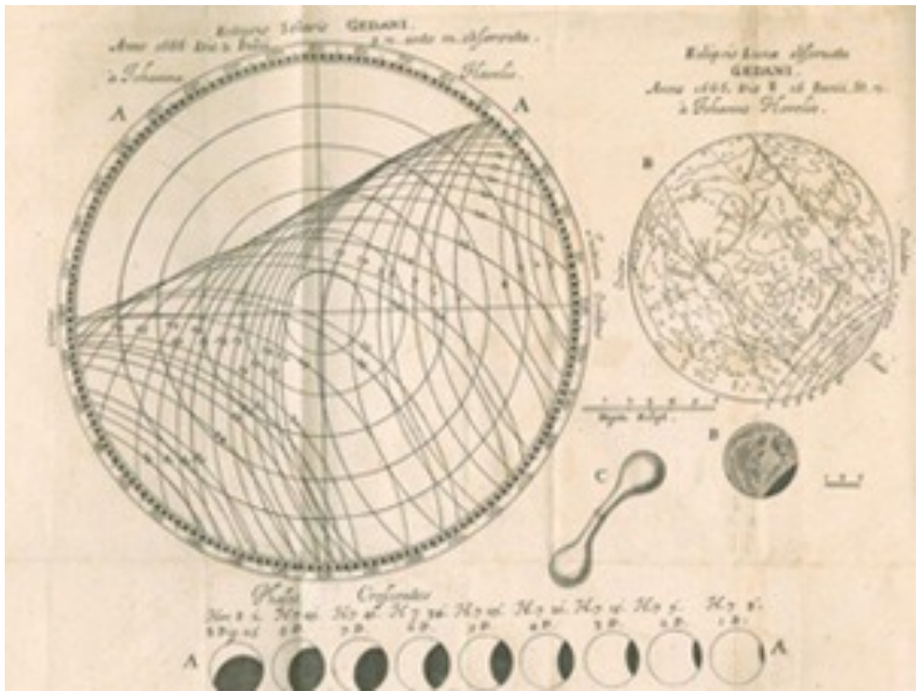


Summary

- In the last decade, *data publishing* emerges as a field in science communication
- Dataverse plays a key role leading data publishing and promoting best practices for making data accessible
- Dataverse is expanding to support data provenance, sensitive data and large scale data, while connecting data with publications and analysis.
- The Structural Biology Data Grid project shows how to:
    - make data accessible in a sustainable way
    - follow best practices in data publishing
    - support larger scale data
    - integrate data with analysis

“The 21st of April, 1665, about eight in the morning, I bored a hole in the body of a fair and large Birch, and put in a Cork with a Quill in the middle; after a Moment or two it [a sap] began to drop, but yet very softly: Some three Hours after I returned¸ and it had filled a Pint Glass, and then it dropped exceeding fast, viz. every Pulse a Drop: This Liquor is not unpleasant to the Taste, and not thick or troubled; yet it looks as though some few drops of Milk were split in a Bason of Fountain Water.”

(Lister, 1697)



Schematic of solar and lunar eclipses in a 1665 paper by Hevelius

(Philosophical Transactions, Vol 1)

Source: Gross Harmon, Reidy, *Communicating Science,* 2002