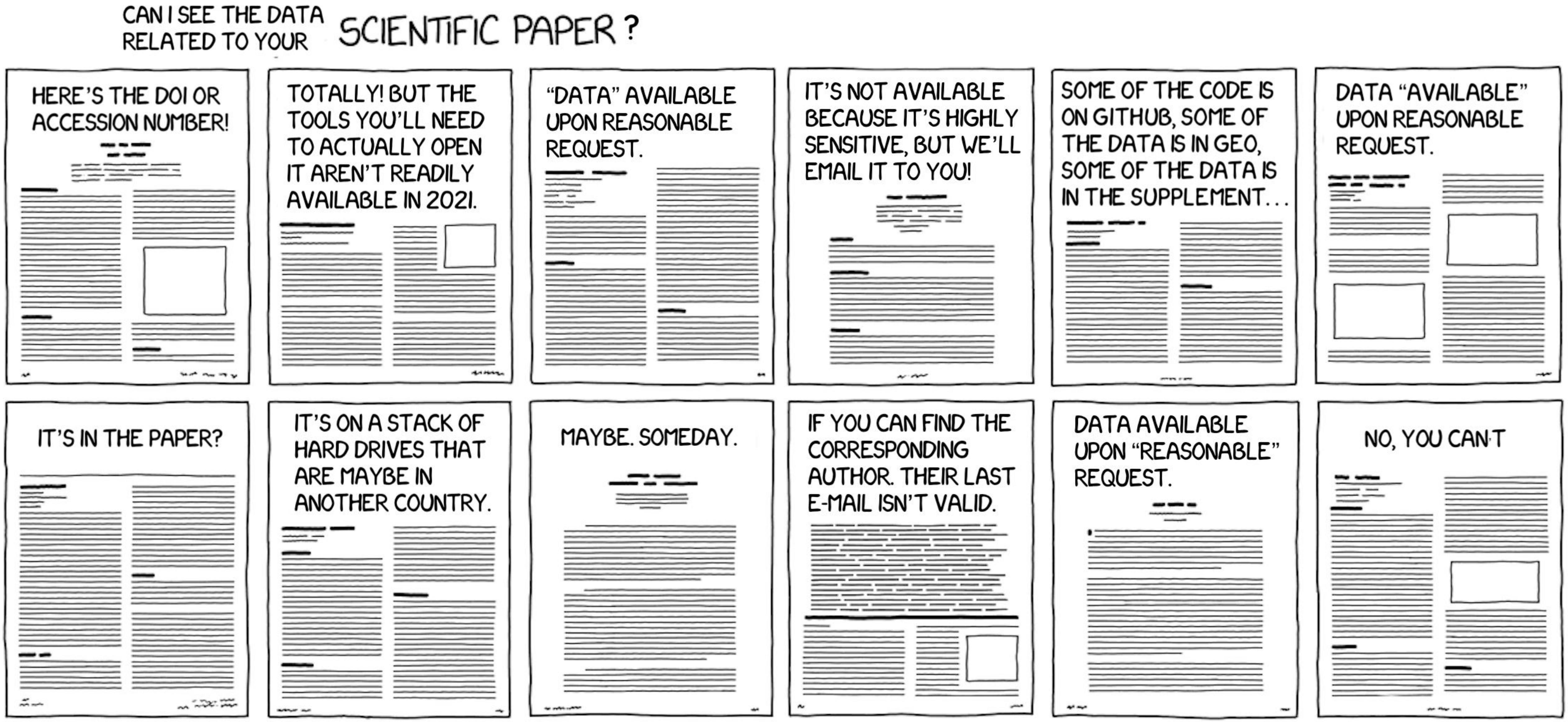
csv,conf,6 | May 5th 2021

Identifying the who, what, and (sometimes) where of research data sharing at an academic institution

John Borghi, PhD
Manager, Research and Instruction
Lane Medical Library
Stanford University


CAN I SEE THE DATA RELATED TO YOUR SCIENTIFIC PAPER?
HERE'S THE DOI OR ACCESSION NUMBER!
TOTALLY! BUT THE TOOLS YOU'LL NEED TO ACTUALLY OPEN IT AREN'T READILY AVAILABLE IN 2021.
"DATA" AVAILABLE UPON REASONABLE REQUEST.
IT'S NOT AVAILABLE BECAUSE IT'S HIGHLY SENSITIVE, BUT WE'LL EMAIL IT TO YOU!
SOME OF THE CODE IS ON GITHUB, SOME OF THE DATA IS IN GEO, SOME OF THE DATA IS IN THE SUPPLEMENT...
DATA "AVAILABLE" UPON REASONABLE REQUEST.
IT'S IN THE PAPER?
IT'S ON A STACK OF HARD DRIVES THAT ARE MAYBE IN ANOTHER COUNTRY.
MAYBE. SOMEDAY.
IF YOU CAN FIND THE CORRESPONDING AUTHOR. THEIR LAST E-MAIL ISN'T VALID.
DATA AVAILABLE UPON "REASONABLE" REQUEST.
NO, YOU CAN'T

Adapted from: https://xkcd.com/2456/

A substantial part of my job is helping researchers at my institution manage and share their data.

A substantial part of my job is helping researchers at my institution manage and share their data.

Research Data: The inputs or outputs required to evaluate, reproduce, or built upon the analyses or conclusions of a given research project. Includes code and documentation.

Data Management: Activities related to data storage, organization, documentation, and communication.

Data Sharing: The release of data for use by others. Doesn’t necessarily refer to making data open.

Who is sharing their data?

A substantial part of my job is helping researchers at my institution manage and share their data.

What kind of data are they sharing?

How are they sharing?

Do they know that I exist?

Artwork by August Isler Borghi, Age 4

One way I’ve looked for researchers and their data is by looking outside the university – at PubMed Central.

PubMed Central

- A repository containing the full text of almost 7 million biomedical and life sciences articles.
- Is distinct from PubMed, the free search engine covering (among other things) the MEDLINE database of references and abstracts.
- Includes articles covered by the NIH Public Access Policy as well as the contents of participating journals.
- Has filters to locate articles with associated data sets, including those with data availability statements.

Data Availability Statements

- Tells the reader how to access the data supporting the results of a given scholarly work.
- May contain direct links to publicly available datasets and/or conditions for accessing more restricted datasets.
- Can provide information about what’s available (and what isn’t).

“All of the data associated with this study have been deposited in [name of repository] at [DOI/Accession number]”

“All data used in this study are available within the article and its supplementary information files.”

“The data used and analyzed in the present study are available from the corresponding author on reasonable request.”

“Data is not publicly available due to participant confidentiality.”

“This manuscript has no associated data.”

“The datasets generated during the current study are not available due to protection of privacy but are available from the corresponding author on reasonable request.”

“All data used in this study are available within the article and its supplementary information files. Other files can be provided by the corresponding author upon reasonable request. The code is available via GitHub [Link to GitHub repository].”

“Most of the raw data are presented in the manuscript or supplement.”

“Some study data is available.”

Limits of this approach

- Incomplete coverage: There are about 10x as many abstracts of papers from Stanford authors in PubMed than there are papers from Stanford authors in PubMed Central with data availability statements.
- Statements are unstructured and may not contain all relevant information.
- Data stated to be available may not actually be available or usable in practice.

((has data avail[filter] OR has data citations[filter]) AND Stanford[AD] AND ("2020/01/01"[PDat] : "2020/12/31"[PDat]))

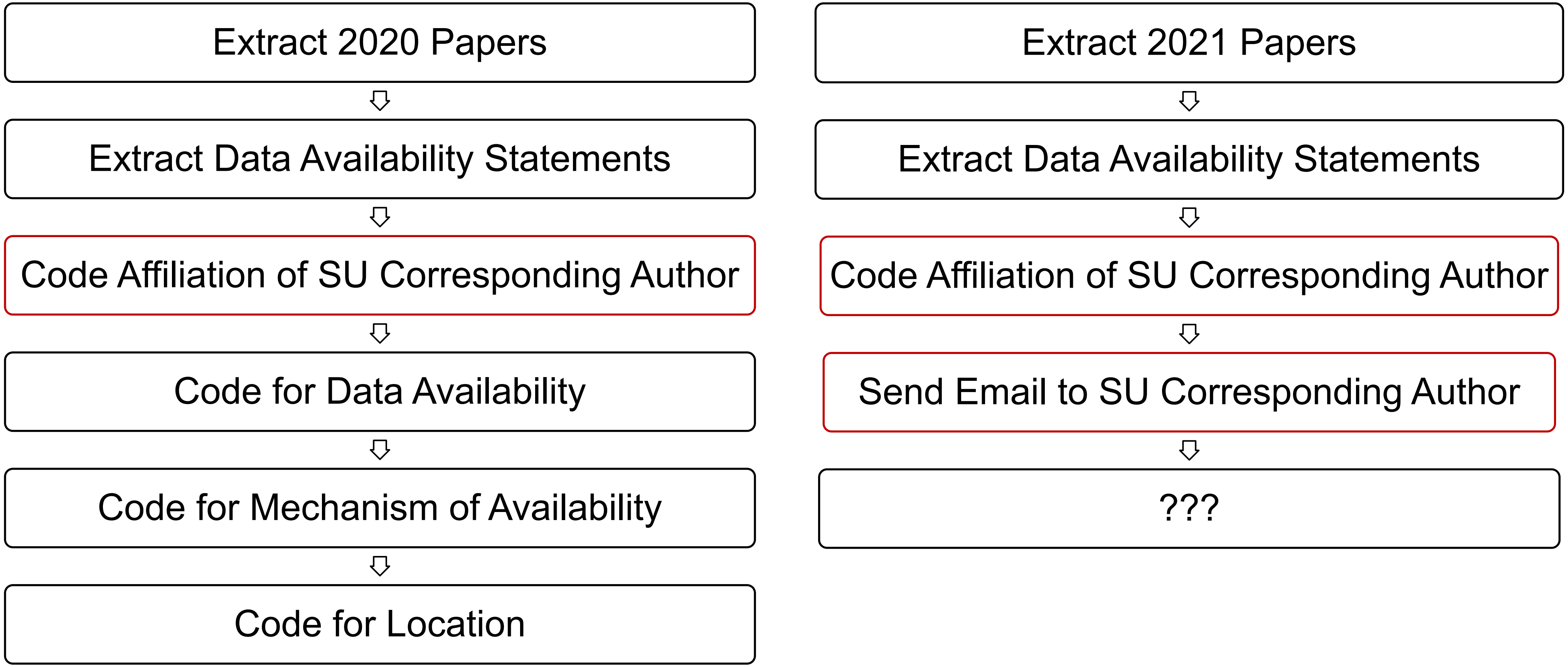
1422 Results

((has data avail[filter] OR has data citations[filter]) AND Stanford[AD] AND ("2021/01/01"[PDat] : "2021/12/31"[PDat]))

540 Results

Results as of May 3rd, 2021 | 12:30 PM

Every week, I check for new results in PubMed Central and then…



Extract 2020 Papers

Extract Data Availability Statements

Code Affiliation of SU Corresponding Author

Code for Data Availability

Code for Mechanism of Availability

Code for Location

Extract 2021 Papers

Extract Data Availability Statements

Code Affiliation of SU Corresponding Author

Send Email to SU Corresponding Author

???

Sample e-mail to a Stanford SOM-affiliated corresponding author

TO: Corresponding Author from Stanford SOM
FROM: John Borghi

SUBJECT: Data Services at Stanford School of Medicine

Hello,

To help us provide research data-related services to the Stanford Medicine community, Lane Medical Library periodically looks into where Stanford researchers are sharing their data, code, and other materials. One of the ways we do this is through looking for papers in PubMed Central with Stanford authors that have data availability statements. For example, your paper [------] was recently deposited into PMC.

I am reaching out to make you aware of some of the data-related resources available to you through Lane Library. Whenever possible, we recommend sharing data through a data repository rather than “by request” or as a supplementary material. Similarly, we recommending archiving code associated with a publication in a repository like Zenodo rather sharing a link to a Github repo.

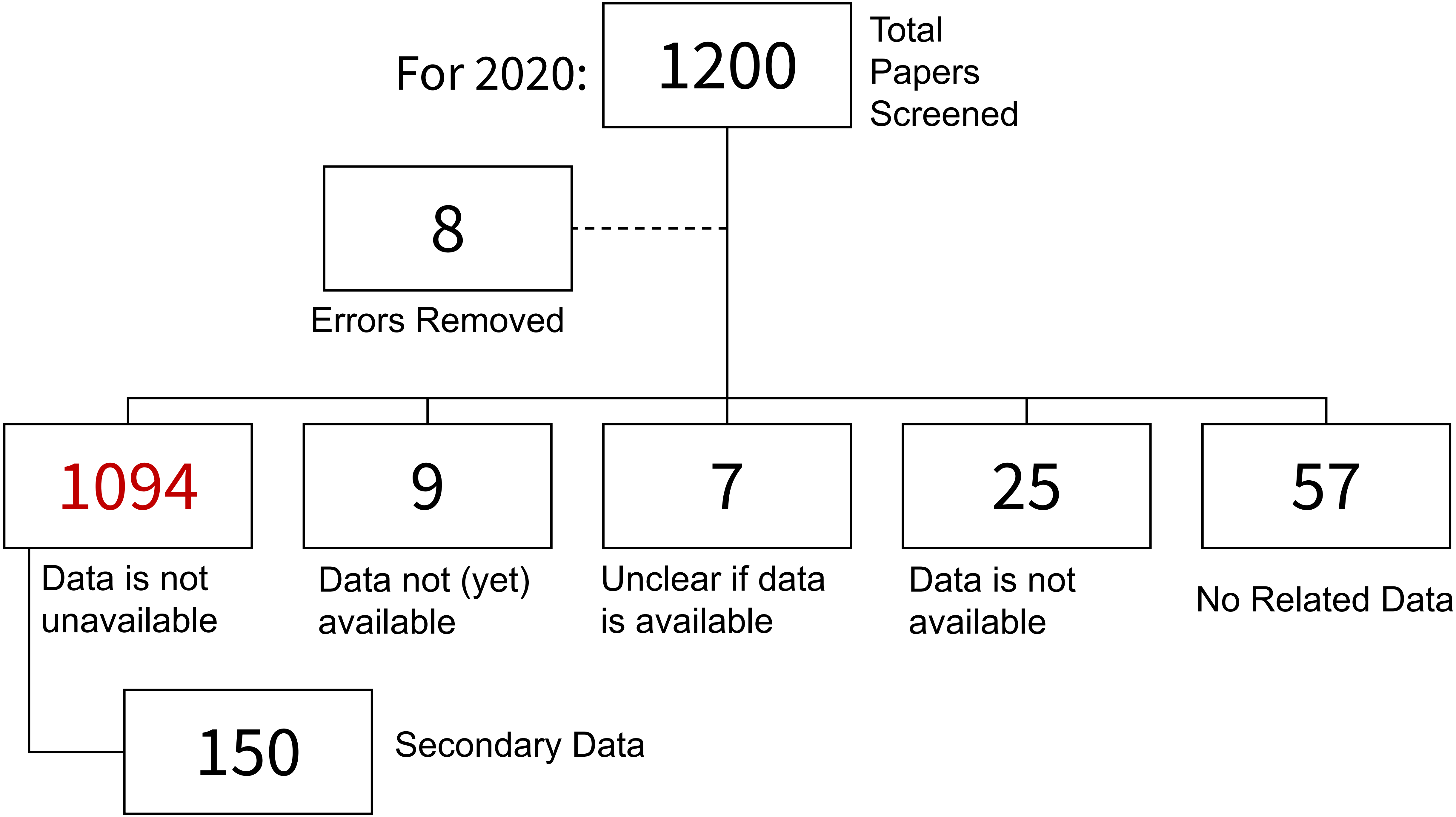
Through Lane, Stanford is a member of the Dryad data repository. This means that Stanford researchers are able to publish and share their data and code through Dryad free of charge. Data deposited into Dryad are automatically given a DOI, lightly curated to ensure files can be opened and do not inadvertently include protected health information, and are preserved over the long term. Code uploaded to Dryad is automatically deposited into Zenodo.

In addition to our Dryad partnership, Lane Library currently maintains this page as a source of information about data management and sharing-related best practices and holds free workshops on these topics (see our classes and events page for our upcoming schedule or request an on-demand version for your group). We are also available for 1 on 1 consultations to help you meet data-related requirements from funding agencies, scholarly publishers, and other stakeholders.

In general, our goal is to meet your where you are and help you get to where you need to be. We’ve partnered with Dryad because we believe it is an excellent general purpose data repository, but we want to help you find the repository that works best for you and your data. We also know that data-related expectations and requirements are constantly evolving, including the new policy from NIH which will come into effect in 2023, so we strive be dynamic and proactive in terms of how we offer our services.

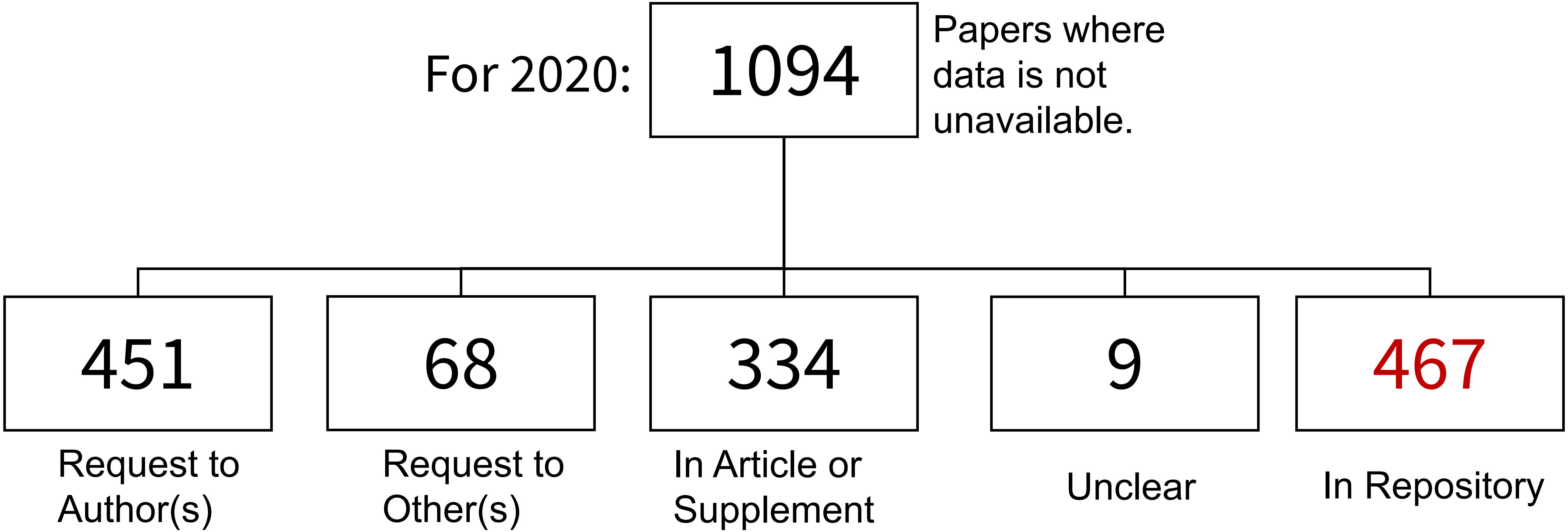
Thank you,
John Borghi

For 2020: 1200 Total Papers Screened


For 2020:
1200
Total Papers Screened
8
Errors Removed
1094
Data is not unavailable
9
Data not (yet) available
7
Unclear if data is available
25
Data is not available
57
No Related Data
150
Secondary Data

For 2020: 1094

Papers where data is not unavailable.



For 2020: 1094 Papers where data is not unavailable.

451 Request to Author(s)

68 Request to Other(s)

334 In Article or Supplement

9 Unclear

467 In Repository

How to Share Data (An hour workshop condensed to one slide)

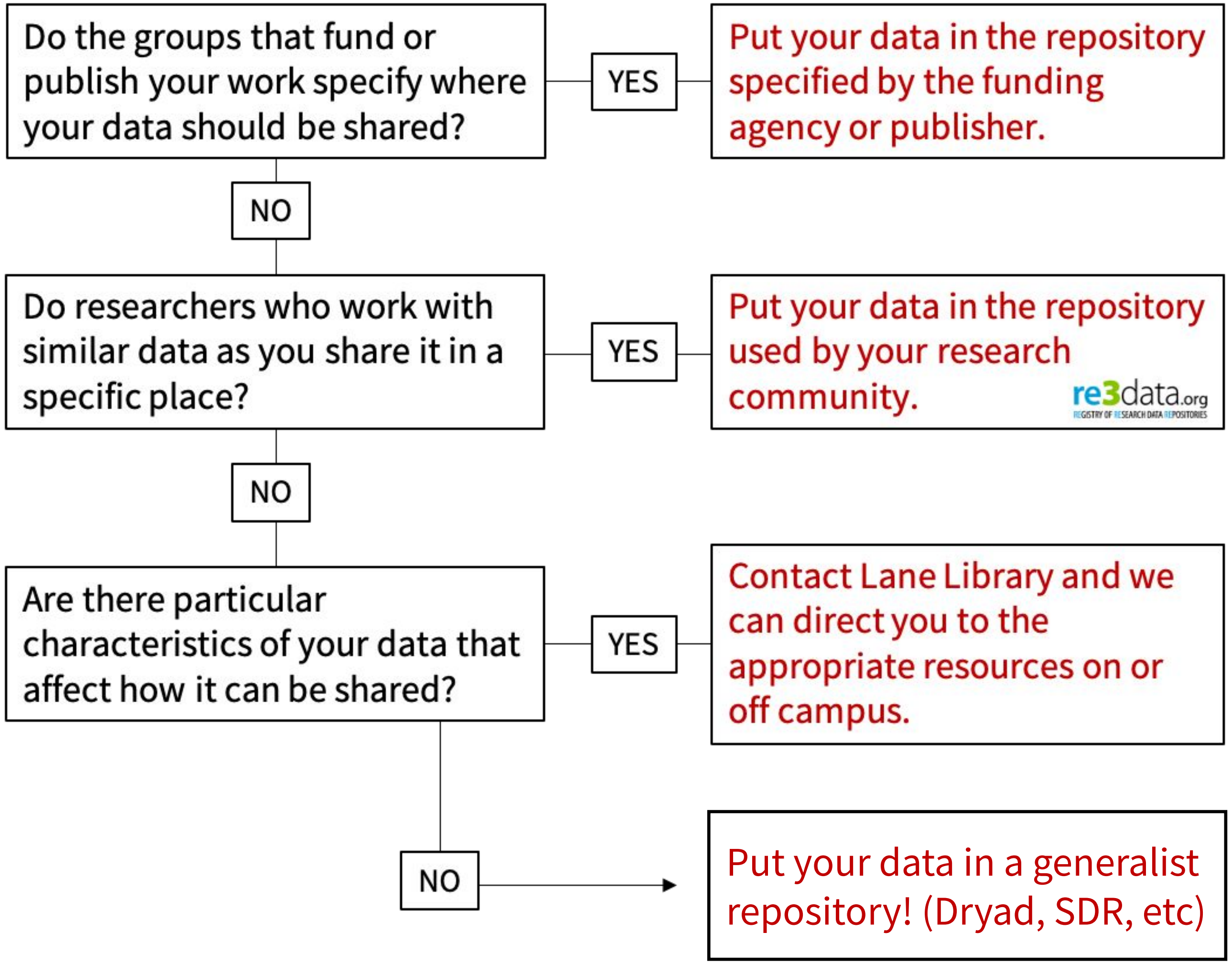
Sharing “upon request”

- When data is shared “upon request”, the requester must contact a member of the research team and a team member must have the data on-hand.
- As contact information changes, team members move on, and data is archive requesting data and responding to requests become more difficult.

Sharing data as a supplementary material

- Links between supplementary materials and the articles they are associated with may break down.
- Whenever feasible, upload data into a repository that is designed to preserve and make data then link to and/or cite that dataset in your manuscript.

Choosing a repository



Do the groups that fund or publish your work specify where your data should be shared?

YES

Put your data in the repository specified by the funding agency or publisher.

NO

Do researchers who work with similar data as you share it in a specific place?

YES

Put your data in the repository used by your research community.

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

NO

Are there particular characteristics of your data that affect how it can be shared?

YES

Contact Lane Library and we can direct you to the appropriate resources on or off campus.

NO

Put your data in a generalist repository! (Dryad, SDR, etc)

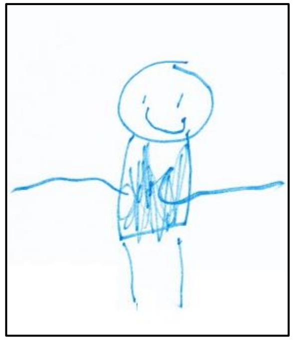
Github	125
Gene Expression Omnibus	103
Website	53
Sequence Read Archive	45
Protein Data Bank	30
Figshare	29
Open Science Framework	25
Dryad	23
BioProject	23
Zenodo	19
Electron Microscopy Data Bank	15
European Genome/Phenome Archive	13
Database of Genotypes and Phenotypes	13
Mendeley Data	12
Genbank	12
Stanford Digital Repository	11
Proteomics Identification Database	11
Flow Repository	10

European Nucleotide Archive	7
SimTK	5
Harvard Dataverse	5
Institutional Repository	4
GTEx	4
ClinVar	4
The Cancer Genome Atlas	3
Software Heritage	3
Open Neuro	3
NIMH Data Archive	3
NCI Genomic Data Commons Data Portal	3
EMBL-EBI	3
Unclear	2
Protein Xchange	2
NIA Genetics of Alzheimer’s Disease Data Storage Site	2
Immport	2
Array Express	2

Github	125
Gene Expression Omnibus	103
Website	53
Sequence Read Archive	45
Protein Data Bank	30
Figshare	29
Open Science Framework	25
Dryad	23
BioProject	23
Zenodo	19
Electron Microscopy Data Bank	15
European Genome/Phenome Archive	13
Database of Genotypes and Phenotypes	13
Mendeley Data	12
Genbank	12
Stanford Digital Repository	11
Proteomics Identification Database	11
Flow Repository	10
European Nucleotide Archive	7
SimTK	5
Harvard Dataverse	5
Institutional Repository	4
GTEx	4
ClinVar	4
The Cancer Genome Atlas	3
Software Heritage	3
Open Neuro	3
NIMH Data Archive	3
NCI Genomic Data Commons Data Portal	3
EMBL-EBI	3
Unclear	2
Protein Xchange	2
NIA Genetics of Alzheimer’s Disease Data Storage Site	2
Immport	2
Array Express	2

Github	125
Gene Expression Omnibus	103
Website	53
Sequence Read Archive	45
Protein Data Bank	30
Figshare	29
Open Science Framework	25
Dryad	23
BioProject	23
Zenodo	19
Electron Microscopy Data Bank	15
European Genome/Phenome Archive	13
Database of Genotypes and Phenotypes	13
Mendeley Data	12
Genbank	12
Stanford Digital Repository	11
Proteomics Identification Database	11
Flow Repository	10
European Nucleotide Archive	7
SimTK	5
Harvard Dataverse	5
Institutional Repository	4
GTEx	4
ClinVar	4
The Cancer Genome Atlas	3
Software Heritage	3
Open Neuro	3
NIMH Data Archive	3
NCI Genomic Data Commons Data Portal	3
EMBL-EBI	3
Unclear	2
Protein Xchange	2
NIA Genetics of Alzheimer’s Disease Data Storage Site	2
Immport	2
Array Express	2

Artwork by August Isler Borghi, Age 4



John Borghi
E-Mail: JBorghi@Stanford.edu
Twitter: @JohnBorghi