

Where's the data? A story of data discovery, cleaning, and equality

IASSIST 2023

Alicia Hofelich Mohr, University of Minnesota

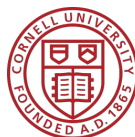
Joel Herndon, Duke University

Cynthia Hudson Vitale, Association of Research Libraries

Realities of Academic Data Sharing (RADS) Study: Metadata Analysis

RADS Institutions:

- Cornell University
- Duke University
- University of Michigan
- University of Minnesota
- Virginia Tech
- Washington University
in St. Louis



Cornell University

Duke
UNIVERSITY



UNIVERSITY OF MINNESOTA



ASSOCIATION
OF RESEARCH
LIBRARIES

DATA
CURATION
NETWORK

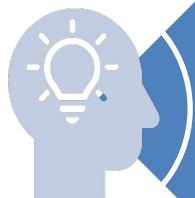


RADS has been funded by NSF EAGER grant
#2135874: Completing the Lifecycle: Developing
Evidence Based Models of Research Data Sharing

RADS Research Questions



Where are funded researchers sharing their data and what is the quality of that metadata?



How are researchers making decisions about why and how to share research data?



What is the cost to the institution to implement federally mandated public access to research data policies?



Realities of Academic Data Sharing research has been generously funded by NSF EAGER grant #2135874: Completing the Lifecycle: Developing Evidence Based Models of Research Data Sharing

ASSOCIATION
OF RESEARCH
LIBRARIES

ACT 1: Where do we look?

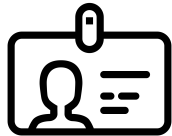
What we set out to do



Identify the location of published data between 2012–2022



Search DOI registries using APIs



Parse affiliation across our six organizations using RORs



Facet results by subject/keywords and funder ID

Done, easy, right?

The best laid plans...



Identify the location of published data between 2012–2022



APIs varied in accessibility & speed → needed to combine approaches

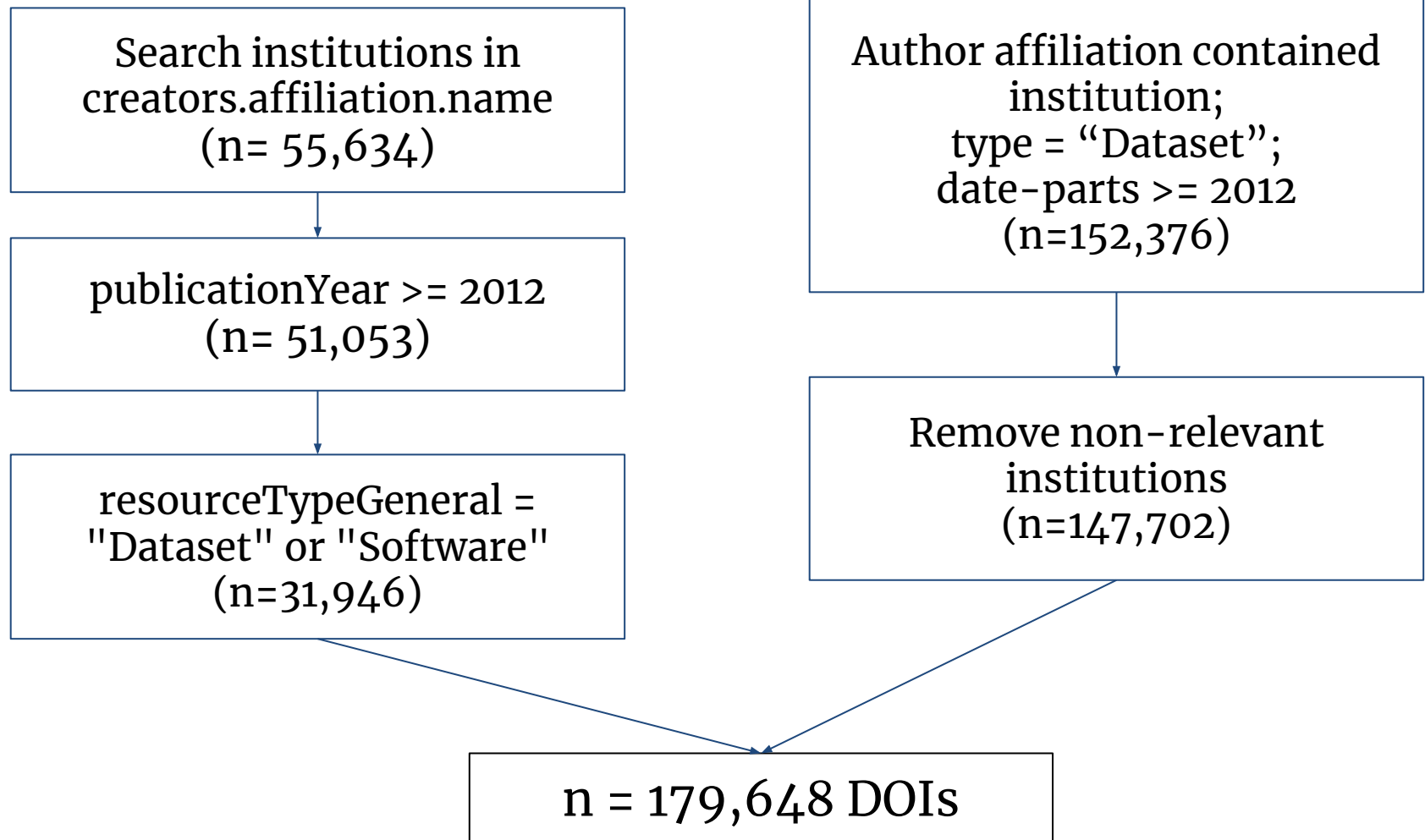


No widespread use of RORs → instead used text search for institution name



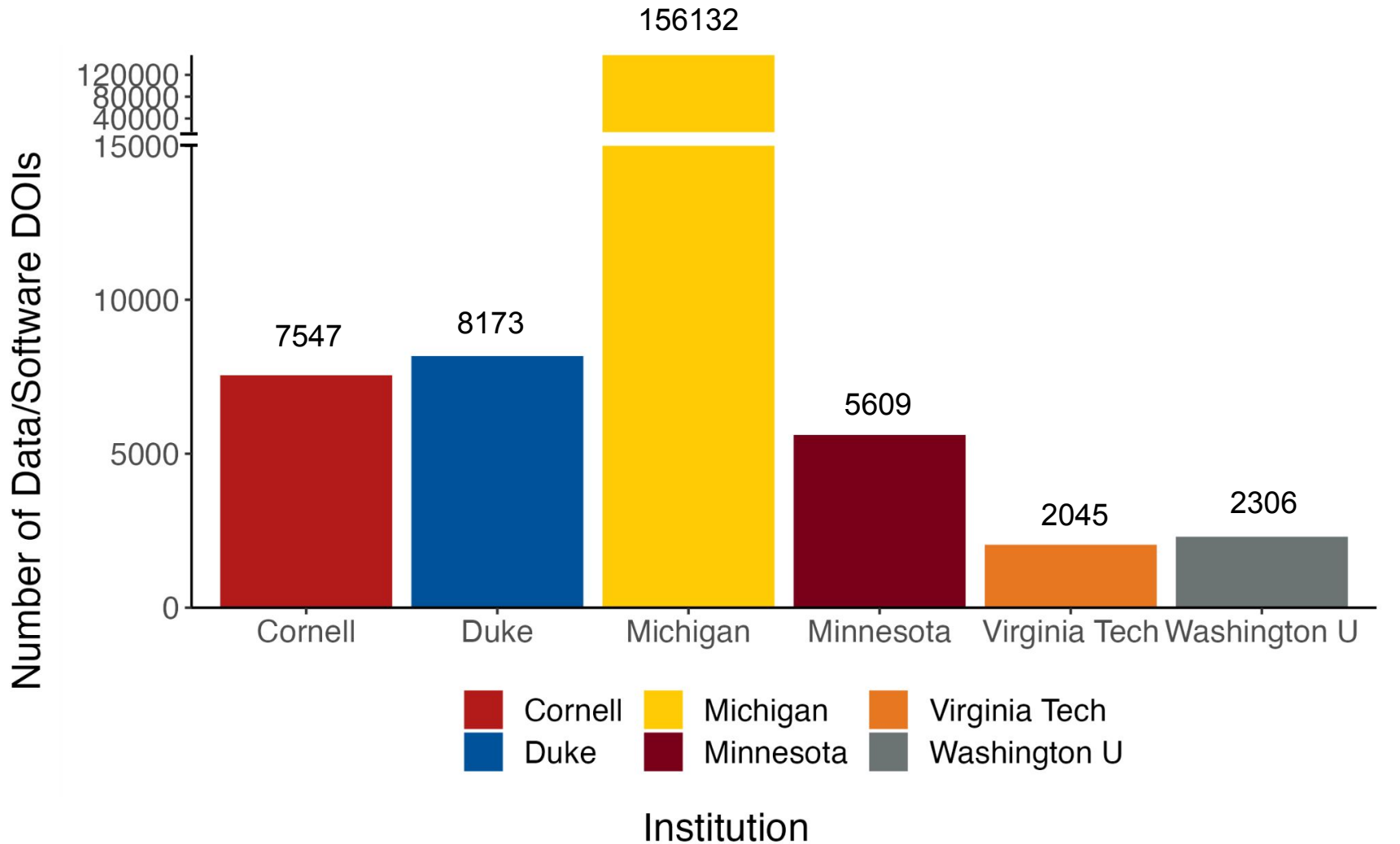
Limited & inconsistent use of funder fields and keywords → removed plans for this facet

The Search



ACT 2: What's here?

Finding all the data



*Michigan cut for scaling

Well, almost...

We knew researchers were publishing data in our institutional repositories...

	Michigan	Minnesota	Cornell	Virginia Tech	WashU	Duke
Institutional Data Repository Records	645	692	174	333	95	225

But we didn't find them

Institutional Repository DOIs	1	90	34	111	16	0
-------------------------------	---	----	----	-----	----	---

Affiliation Woes

- None of us consistently entered "affiliation" metadata in DataCite/CrossRef.
- Solution: Search for our repositories



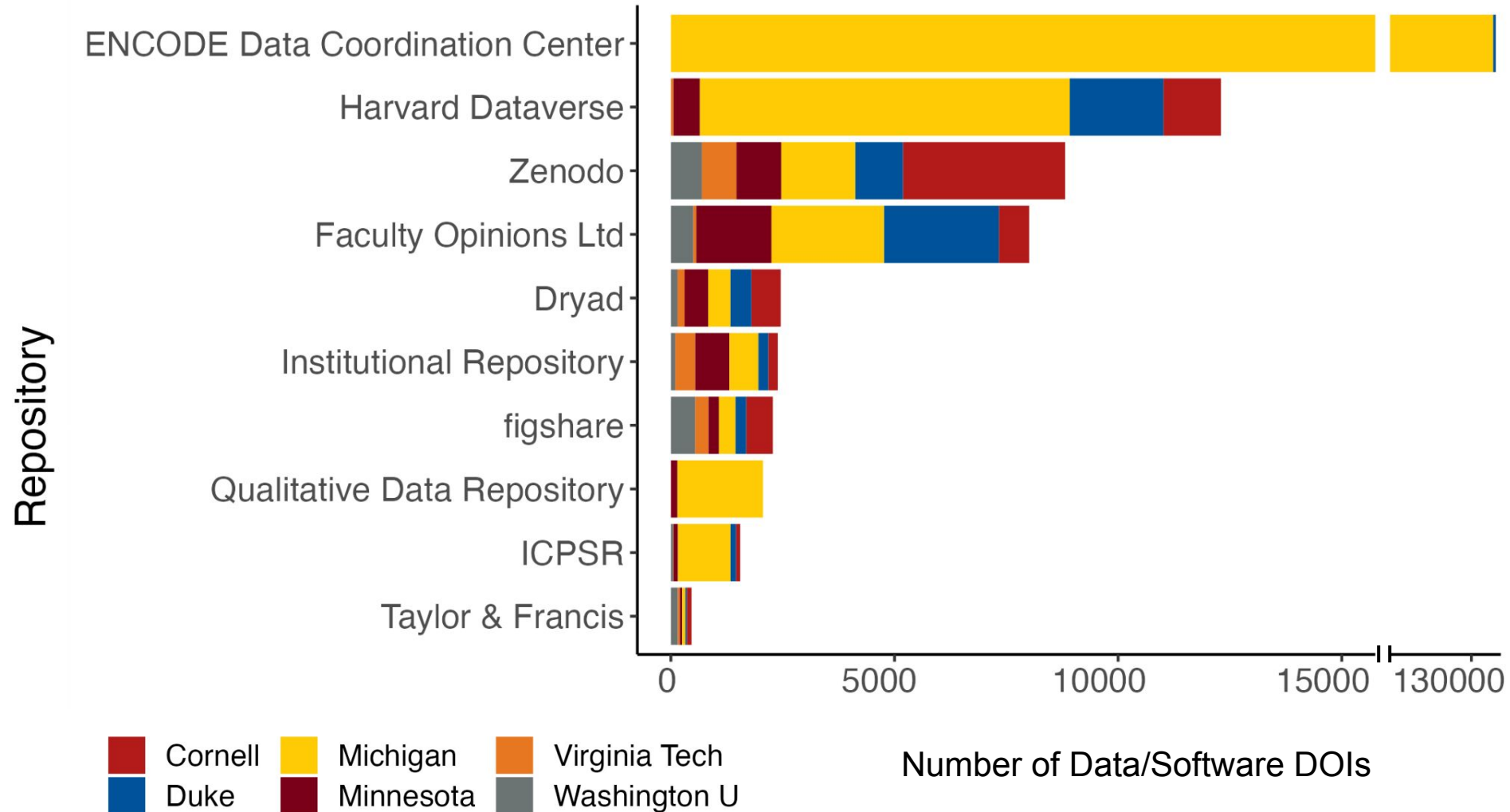
Search five institutional
repositories in publisher;
publicationYear >= 2012;
generalResourceType =
dataset or software
(n = 1,939)



Search Duke's member prefix
for year >=2012;
type = data
(n = 225)

Additional n = 2,164 DOIs

Now, where's the data (Top 10)?



A closer look...

Study DOI

HARVARD
Dataverse

Harvard Dataverse >

A Dataset of Robot Bite Acquisition Trials on Solid Food Using Different Manipulation Strategies

Version 1.1

 Feng, Ryan; Kim, Youngsun; Lee, Gilwoo; Gordon, Ethan Kroll; Schmittle, Matt; Kumar, Shivaum; Bhattacharjee, Tapomayukh; Srinivasa, Siddhartha S. 2019. "A Dataset of Robot Bite Acquisition Trials on Solid Food Using Different Manipulation Strategies", <https://doi.org/10.7910/DVN/8A1XO3>, Harvard Dataverse, V1, UNF:6:w3unK4j1I5QgMeZeYuw9og== [file] [INFO]

[Cite Dataset](#) ▾ [Learn about Data Citation Standards.](#)

[Access Dataset](#) ▾

[Contact Owner](#) [Share](#)

Filter by
File Type: All ▾ Access: All ▾

1 to 10 of 2,493 Files [Download](#) ▾

	apple-angle-0-trial-0.zip vertical_skewer_wall/ ZIP Archive - 1.3 MB	 
---	--	---



HARVARD
Dataverse

Harvard Dataverse >
[A Dataset of Robot Bite Acquisition Trials on Solid Food Using Different Manipulation Strategies](#) >

apple-angle-0-trial-0.zip

This file is part of "A Dataset of Robot Bite Acquisition Trials on Solid Food Using Different Manipulation Strategies".

Version 1.1

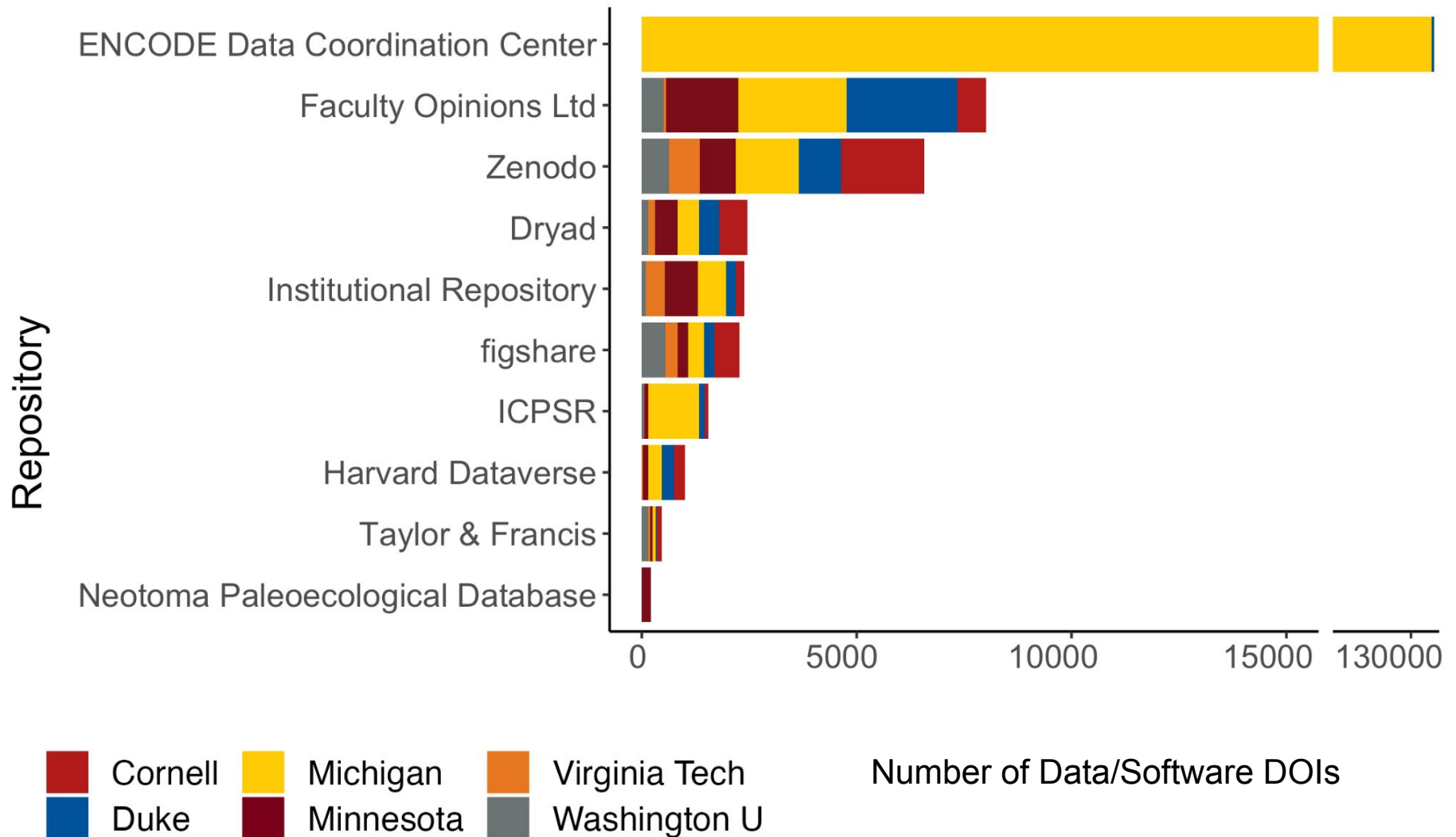
File Citation

Feng, Ryan; Kim, Youngsun; Lee, Gilwoo; Gordon, Ethan Kroll; Schmittle, Matt; Kumar, Shivaum; Bhattacharjee, Tapomayukh; Srinivasa, Siddhartha S. 2019, "apple-angle-0-trial-0.zip", *A Dataset of Robot Bite Acquisition Trials on Solid Food Using Different Manipulation Strategies*, <https://doi.org/10.7910/DVN/8A1XO3/SAHOUD>, Harvard Dataverse, V1

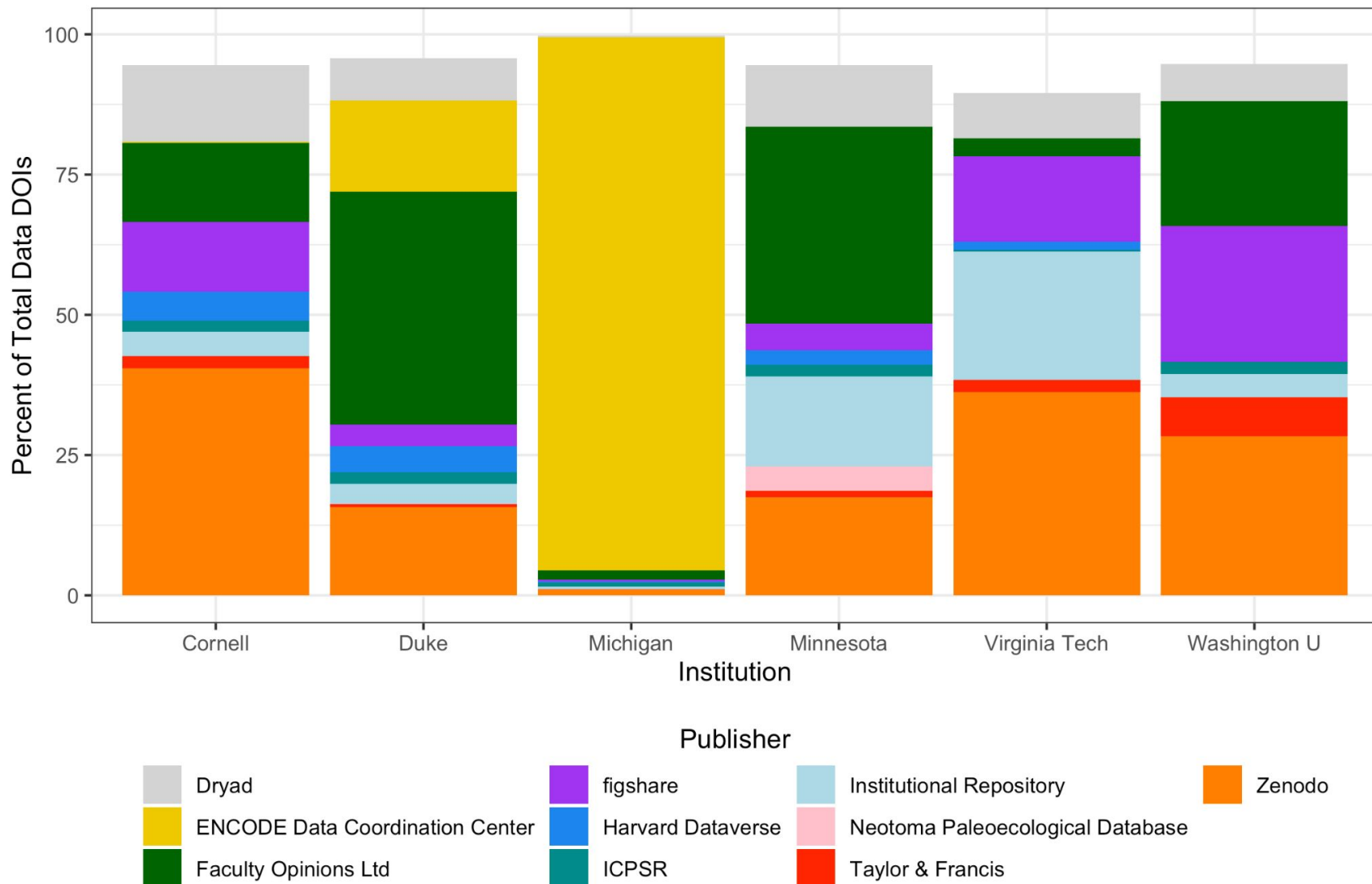
[Cite Data File](#) ▾ [Learn about Data Citation Standards.](#)

AND File DOIs...
(a lot of file DOIs)

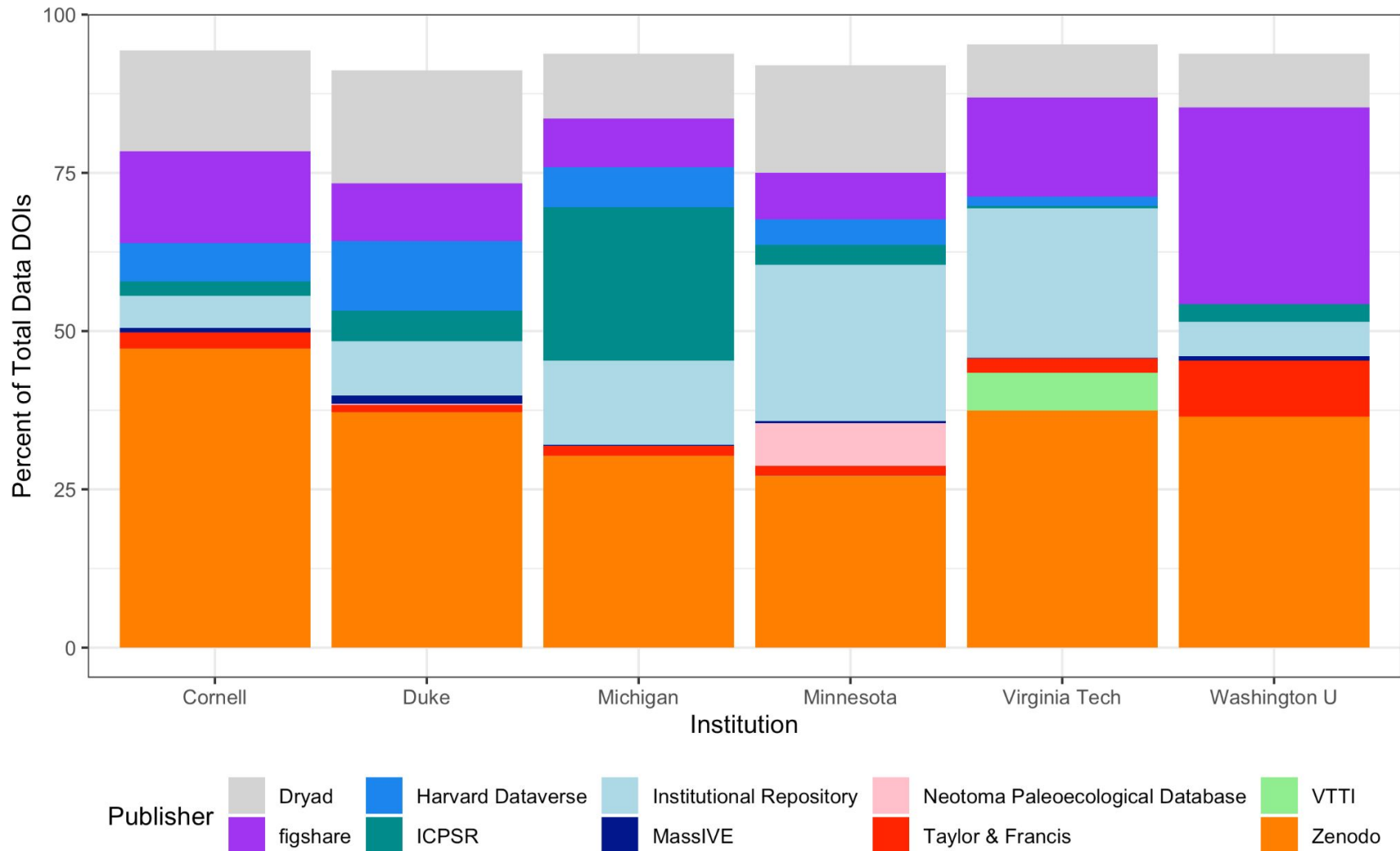
Ok... now where's the data (Top 10)?



Distribution of Publisher by Institution



Without ENCODE and Faculty LTD



ACT 3: What's missing?

Assumptions

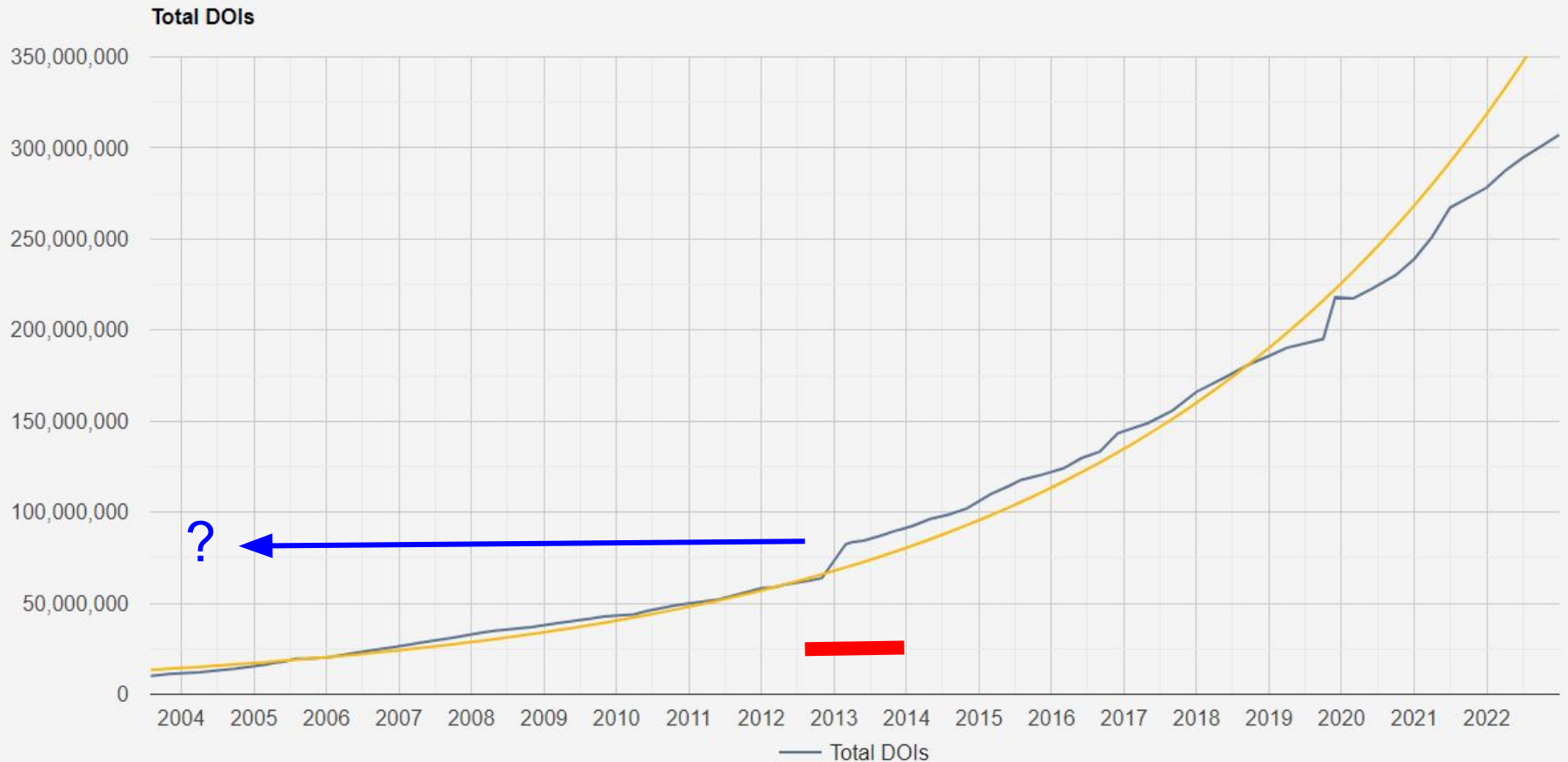
DOI as THE data/software identifier – but is this true?



DOI Growth Over Time

TOTAL DOIS

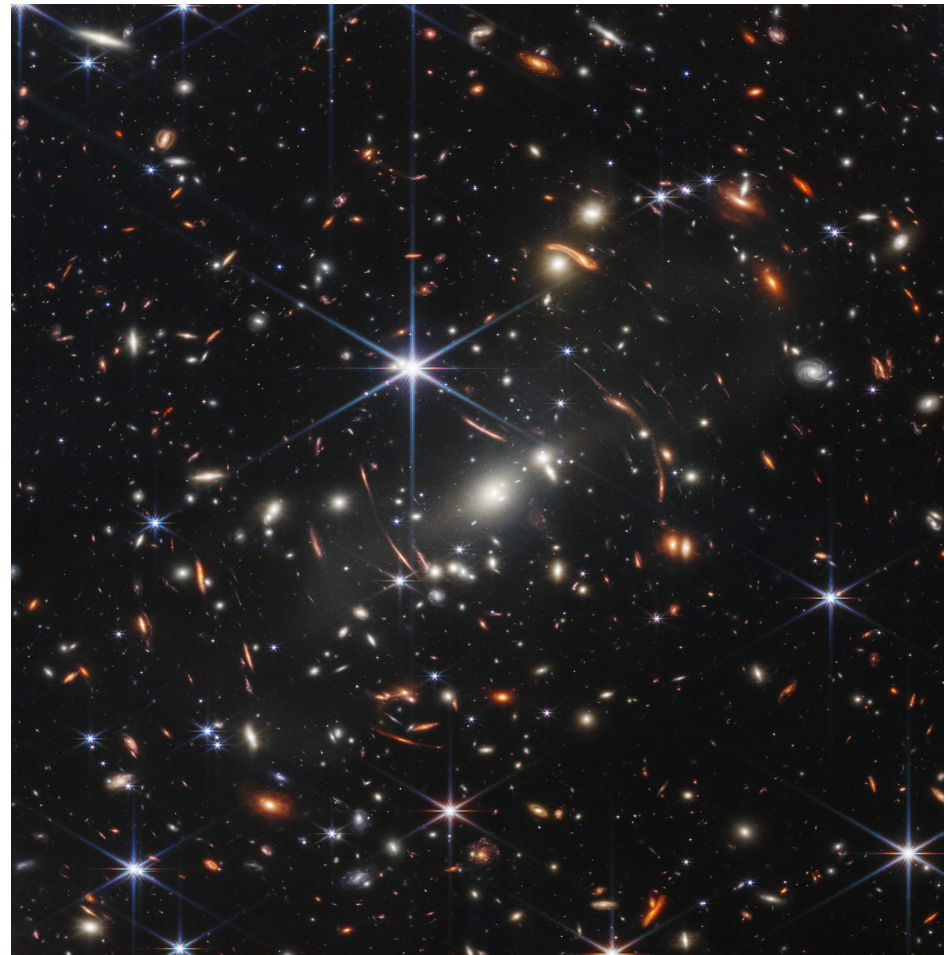
The cumulative number of DOI registered in the DOI System over time; the yellow line shows the moving average



<https://www.doi.org/the-identifier/what-is-a-doi/>

but... the data sharing space is vast...

- Handles, ARKs, other local identifiers for data
- Accession numbers (medical fields)
- No registered identifiers at all (including some linked data)

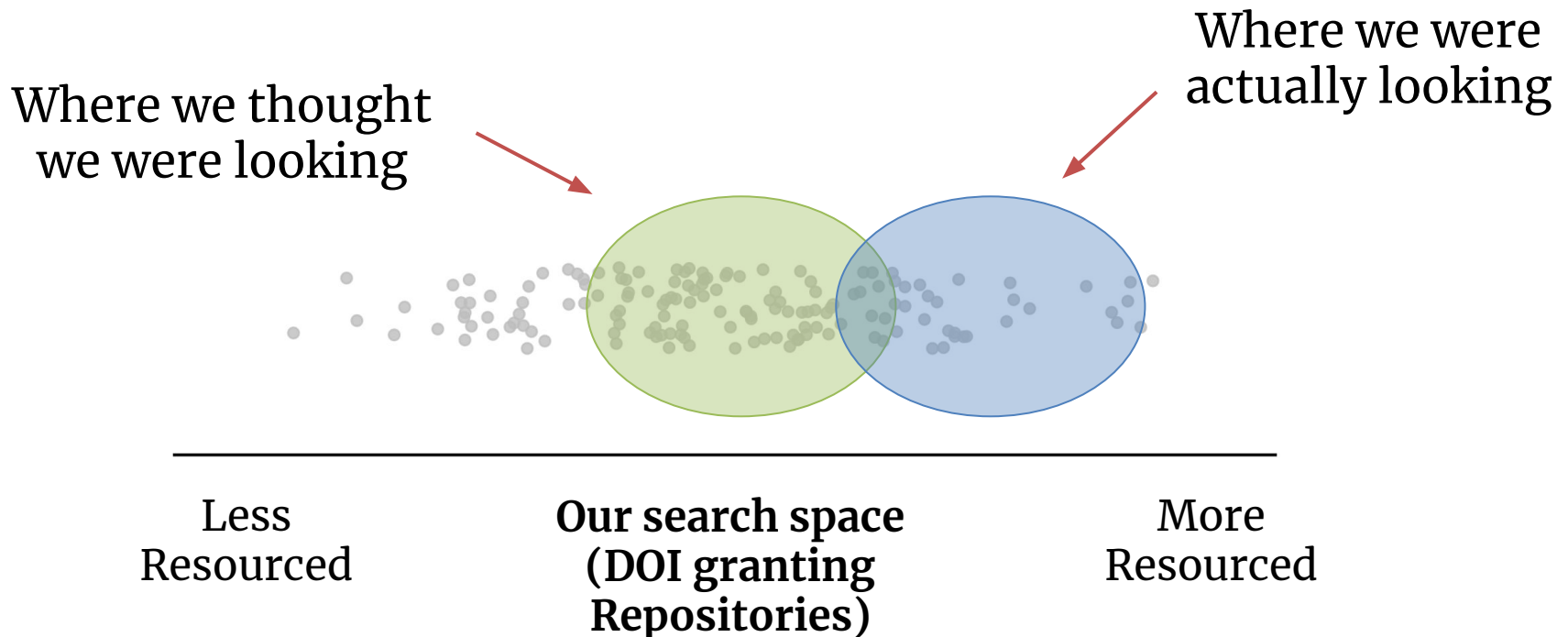


... and (even) DOI's present assessment challenges:

- Multiple registration agencies
- Metadata completeness (required vs “suggested”)
- Granularity (“Itemness”)/Versioning
- Software (sometimes coded as “data”)

But there was bias within our search of DOIs

We are likely “losing” information from smaller, less-resourced repositories.



But there was bias within our search of DOIs

Taking a step back, we really only see a VERY small portion of the data sharing picture

- Metadata findability is highly biased towards well resourced repositories

What we found



Data outside
repositories

Data in repositories

Metadata Findability

A black silhouette of a person pushing a large sphere up a black hill against a blue background. The person is positioned on the left side of the frame, leaning forward and pushing the sphere upwards. The sphere is large and black, and the hill is a solid black shape that curves upwards from the bottom left towards the top left.

The struggle is real!

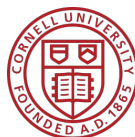
Metadata decisions are ... complicated ...

- Capacity and resource availability
- Balancing description and discovery
- Multiple (evolving) standards
- Correction is time consuming, costly

Realities of Academic Data Sharing (RADS) Study: Metadata Analysis

RADS Institutions:

- Cornell University
- Duke University
- University of Michigan
- University of Minnesota
- Virginia Tech
- Washington University
in St. Louis



Cornell University

Duke
UNIVERSITY



UNIVERSITY OF MINNESOTA



ASSOCIATION
OF RESEARCH
LIBRARIES

DATA
CURATION
NETWORK



RADS has been funded by NSF EAGER grant
#2135874: Completing the Lifecycle: Developing
Evidence Based Models of Research Data Sharing

Thank You!

Contact Us

Find our work on github: [ajhmohr/rads_metadata](https://github.com/ajhmohr/rads_metadata)



**DATA
CURATION
NETWORK**



RADS has been funded by NSF EAGER grant
#2135874: Completing the Lifecycle: Developing
Evidence Based Models of Research Data Sharing