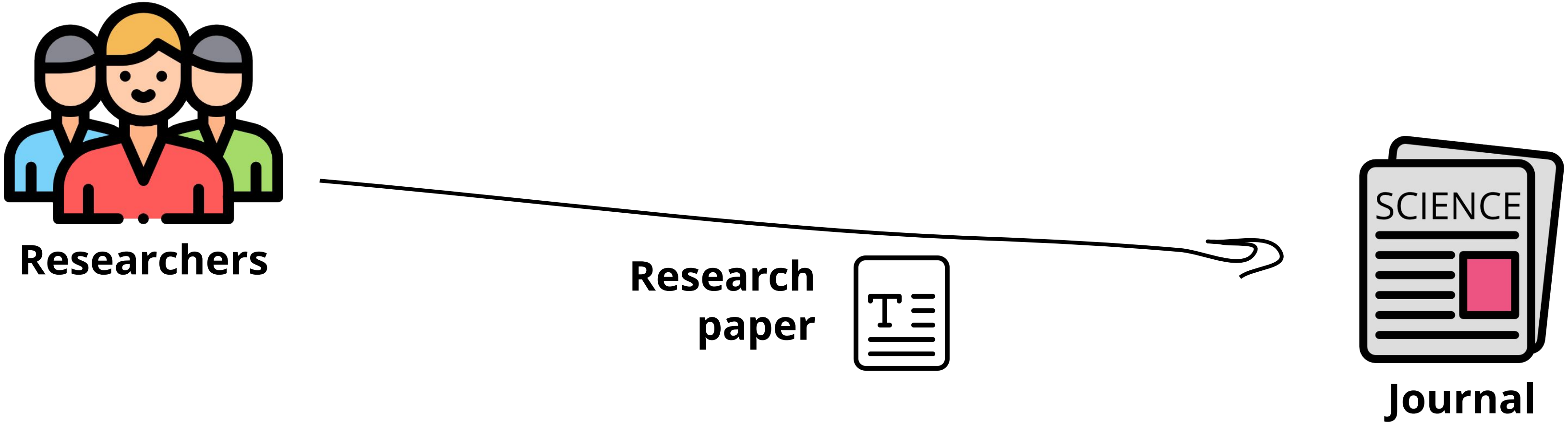
ALFRED P. SLOAN
FOUNDATION

Cluster Analysis of Open Research Data: a Case for Replication Metadata

17th International Digital Curation Conference
June 15, 2022

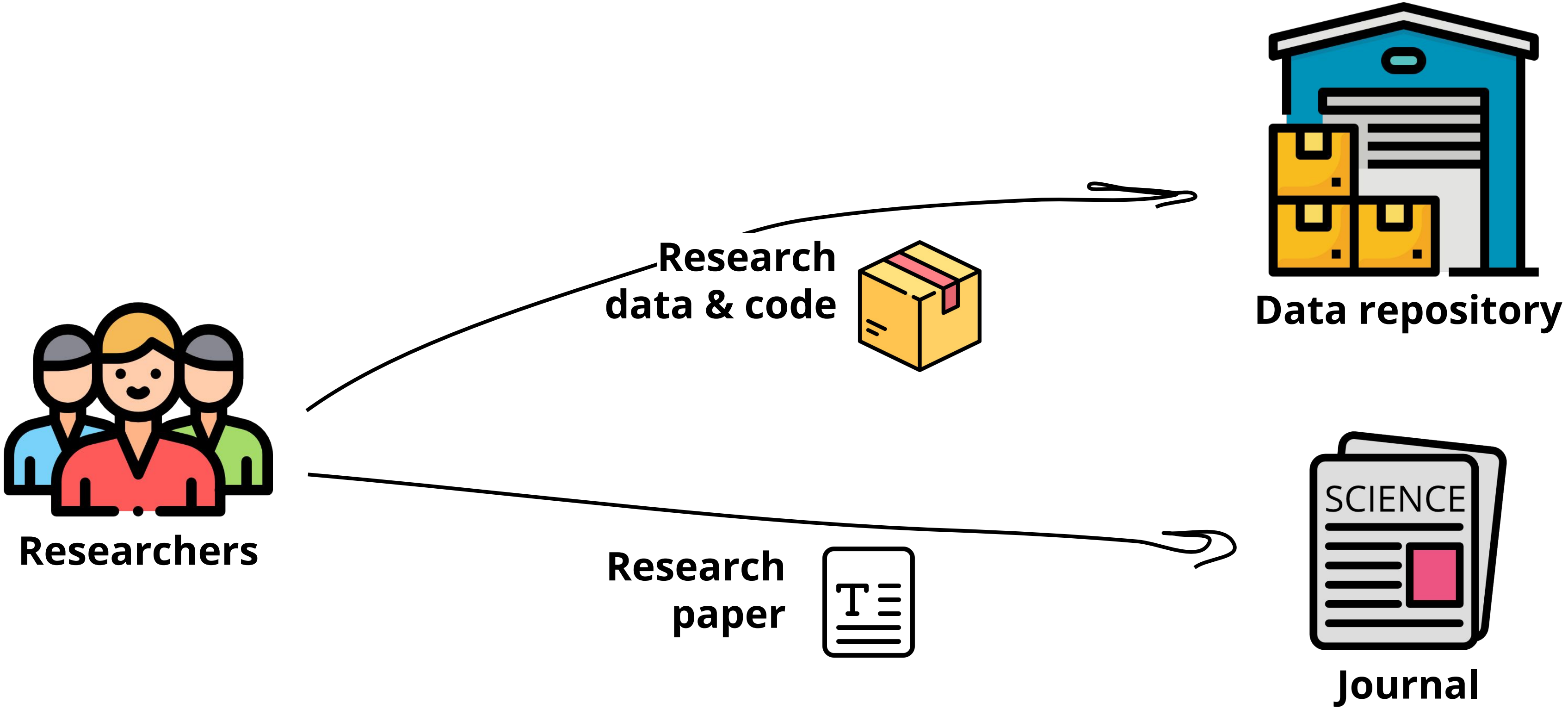
Ana Trisovic, Research Associate
Harvard Biostatistics & the Institute for Quantitative Social Science



Researchers

Research paper

Journal



Research data & code

Data repository

Researchers

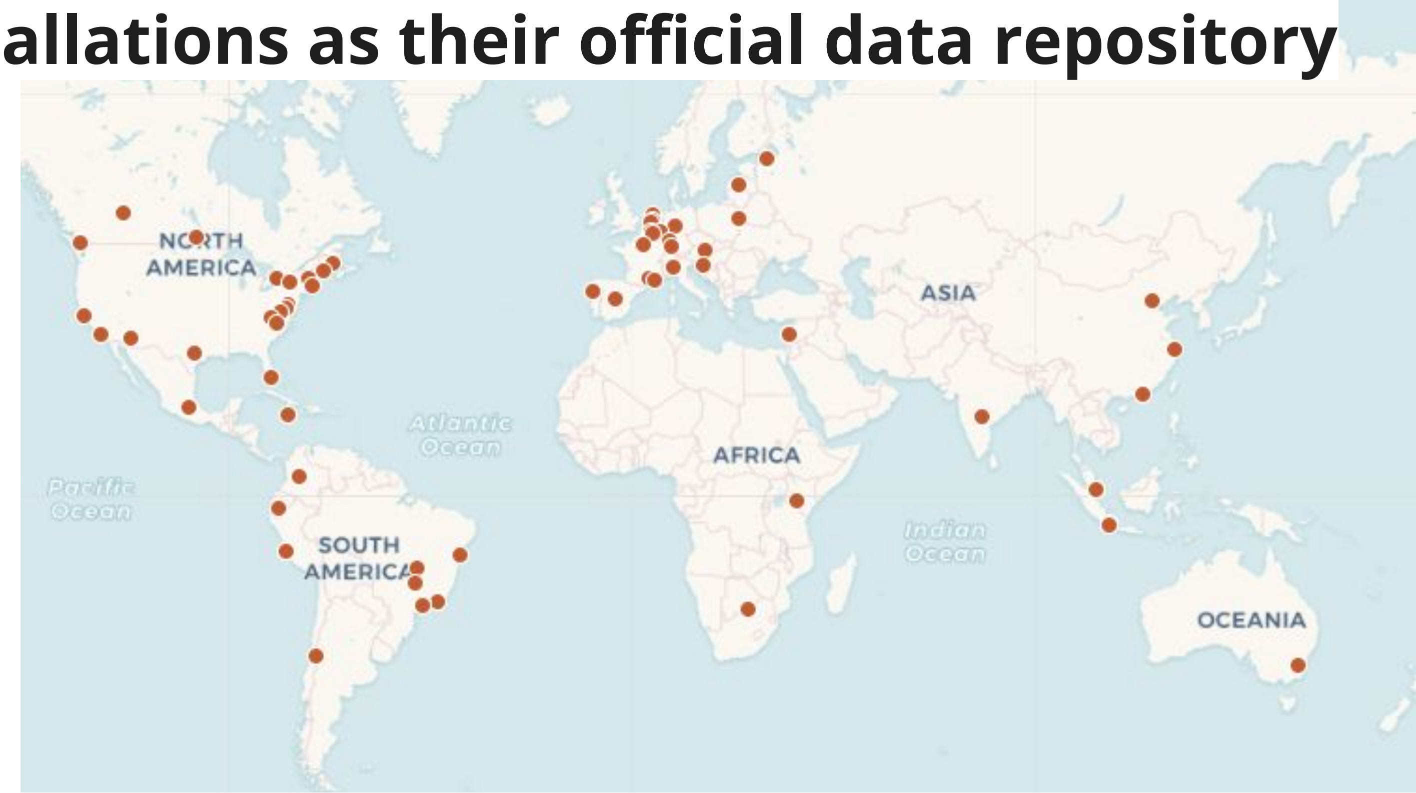
Research paper

Journal

The Dataverse® Project

- A free and open-source software platform to archive, share, and cite research data
- Focus on data sharing and making data available

80 institutions around the globe run Dataverse installations as their official data repository



dataverse.org

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study

Political Analysis Dataverse (Cambridge University Press)

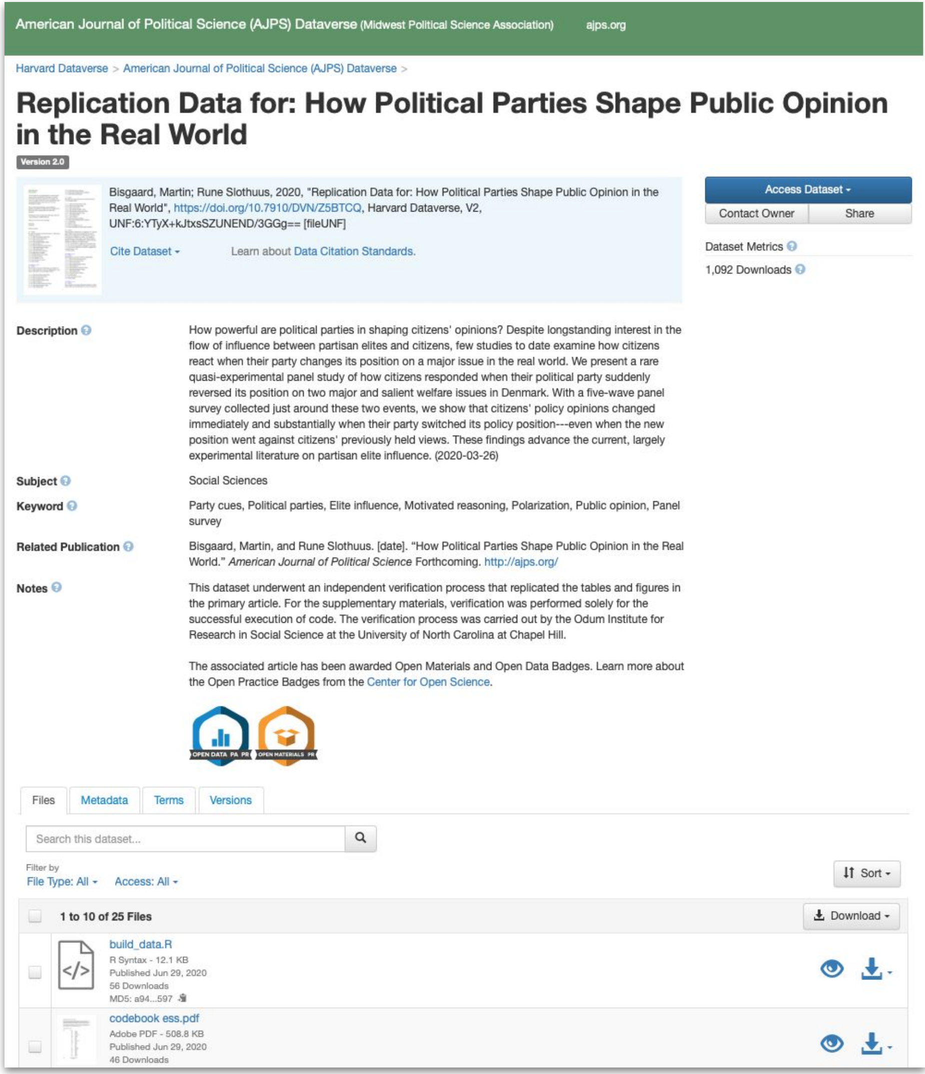
Harvard Dataverse >

Political Analysis is the
methodology, includin

American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse

The *American Journal of Political Science* is committed to significant advances in knowledge and understanding of citizenship, gov
politics, and to the public value of political science research. To find out more about our data integrity policies, please visit our webs



American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >

Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0

Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", https://doi.org/10.7910/DVN/Z5BTCQ, Harvard Dataverse, V2, UNF:6:YTyX+kJtxsSZUNEND/3GGg== [fileUNF]

Learn about Data Citation Standards.

Dataset Metrics

1,092 Downloads

Description How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinions changed immediately and substantially when their party switched its policy position---even when the new position went against citizens' previously held views. These findings advance the current, largely experimental literature on partisan elite influence. (2020-03-26)

Subject Social Sciences

Keyword Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

Related Publication Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. http://ajps.org/

Notes This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

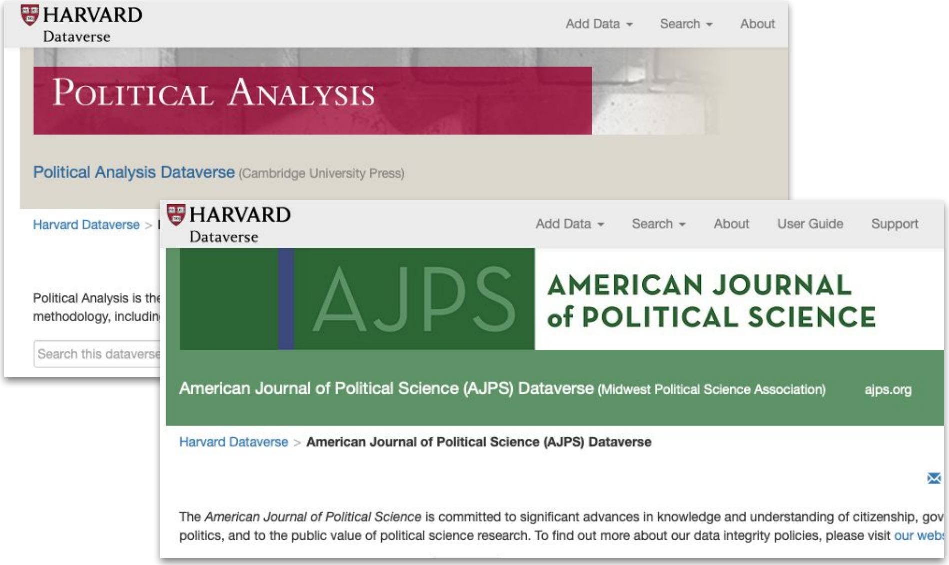
The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the Center for Open Science.

1 to 10 of 25 Files

build_data.R
R Syntax - 12.1 KB
Published Jun 29, 2020
56 Downloads
MD5: a94...597

codebook ess.pdf
Adobe PDF - 508.8 KB
Published Jun 29, 2020
46 Downloads

- Replication dataset - a bundle of data, code and other files needed to reproduce a published study



Political Analysis Dataverse (Cambridge University Press)

Harvard Dataverse >

Political Analysis is the

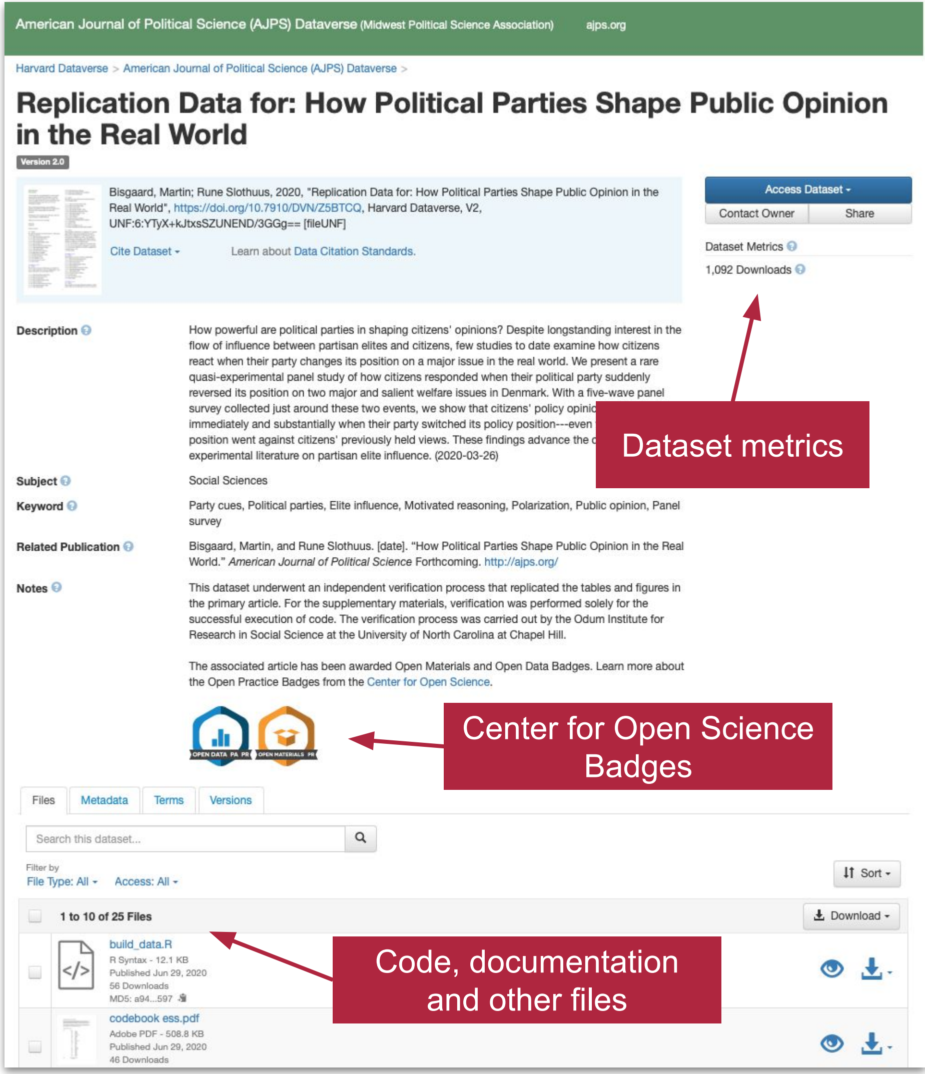
methodology, includin

Search this dataverse

American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse

The *American Journal of Political Science* is committed to significant advances in knowledge and understanding of citizenship, gov politics, and to the public value of political science research. To find out more about our data integrity policies, please visit our webs


American Journal of Political Science (AJPS) Dataverse (Midwest Political Science Association) ajps.org

Harvard Dataverse > American Journal of Political Science (AJPS) Dataverse >

Replication Data for: How Political Parties Shape Public Opinion in the Real World

Version 2.0

Bisgaard, Martin; Rune Slothuus, 2020, "Replication Data for: How Political Parties Shape Public Opinion in the Real World", https://doi.org/10.7910/DVN/Z5BTCQ, Harvard Dataverse, V2, UNF:6:YTyX+kJtxsSZUNEND/3GGg== [fileUNF]

Cite Dataset Learn about Data Citation Standards.

Dataset Metrics

1,092 Downloads

Description: How powerful are political parties in shaping citizens' opinions? Despite longstanding interest in the flow of influence between partisan elites and citizens, few studies to date examine how citizens react when their party changes its position on a major issue in the real world. We present a rare quasi-experimental panel study of how citizens responded when their political party suddenly reversed its position on two major and salient welfare issues in Denmark. With a five-wave panel survey collected just around these two events, we show that citizens' policy opinic immediately and substantially when their party switched its policy position---even position went against citizens' previously held views. These findings advance the c experimental literature on partisan elite influence. (2020-03-26)

Subject: Social Sciences

Keyword: Party cues, Political parties, Elite influence, Motivated reasoning, Polarization, Public opinion, Panel survey

Related Publication: Bisgaard, Martin, and Rune Slothuus. [date]. "How Political Parties Shape Public Opinion in the Real World." *American Journal of Political Science* Forthcoming. http://ajps.org/

Notes: This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the Center for Open Science.

1 to 10 of 25 Files

build_data.R
R Syntax - 12.1 KB
Published Jun 29, 2020
56 Downloads
MD5: a94...597

codebook ess.pdf
Adobe PDF - 508.8 KB
Published Jun 29, 2020
46 Downloads

Dataset metrics

Center for Open Science Badges

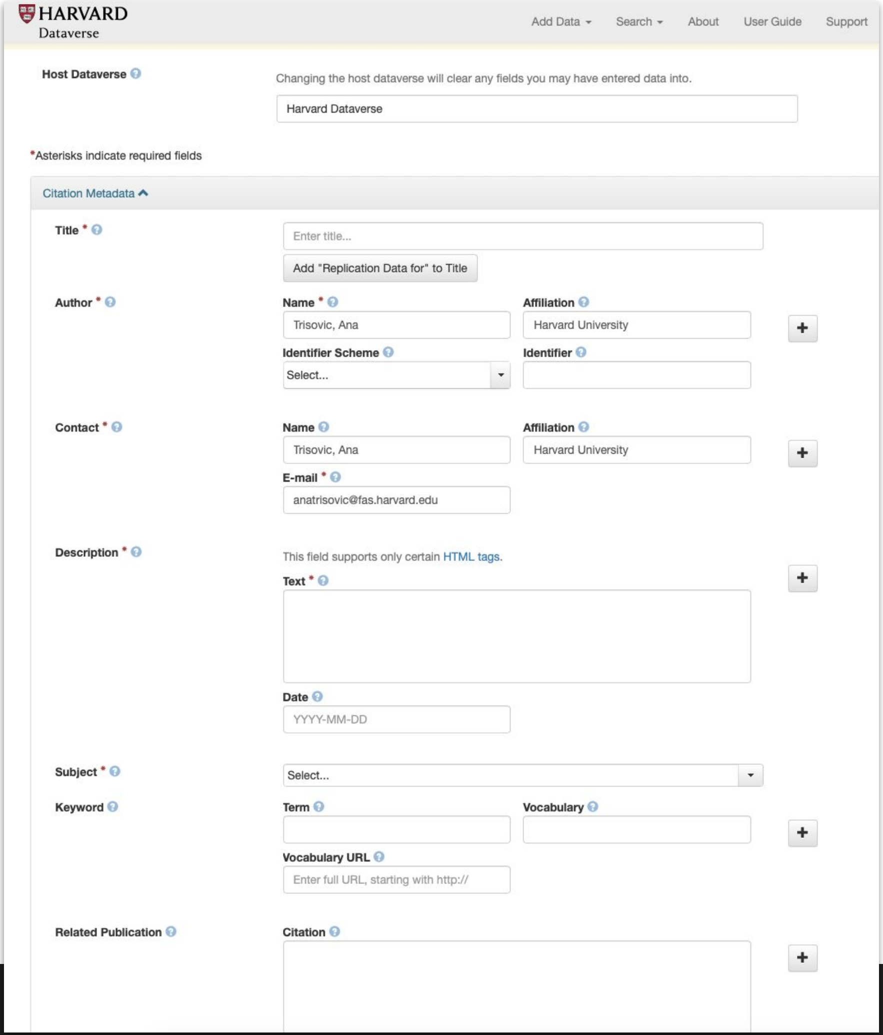
Code, documentation and other files

FAIR principles

Findable	Describe data in metadata, assign DOI Metadata record is shared in data repository
Accessible	Accessible but not necessarily open Standard access protocol
Interoperable	File format open or proprietary Description of data elements
Reusable	License and usage rights Data provenance

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* (2016)

Metadata



Host Dataverse: Changing the host dataverse will clear any fields you may have entered data into.
Harvard Dataverse

*Asterisks indicate required fields

Title *: Enter title...

Author *: Name *: Trisovic, Ana — Affiliation: Harvard University
Identifier Scheme: Select... — Identifier:

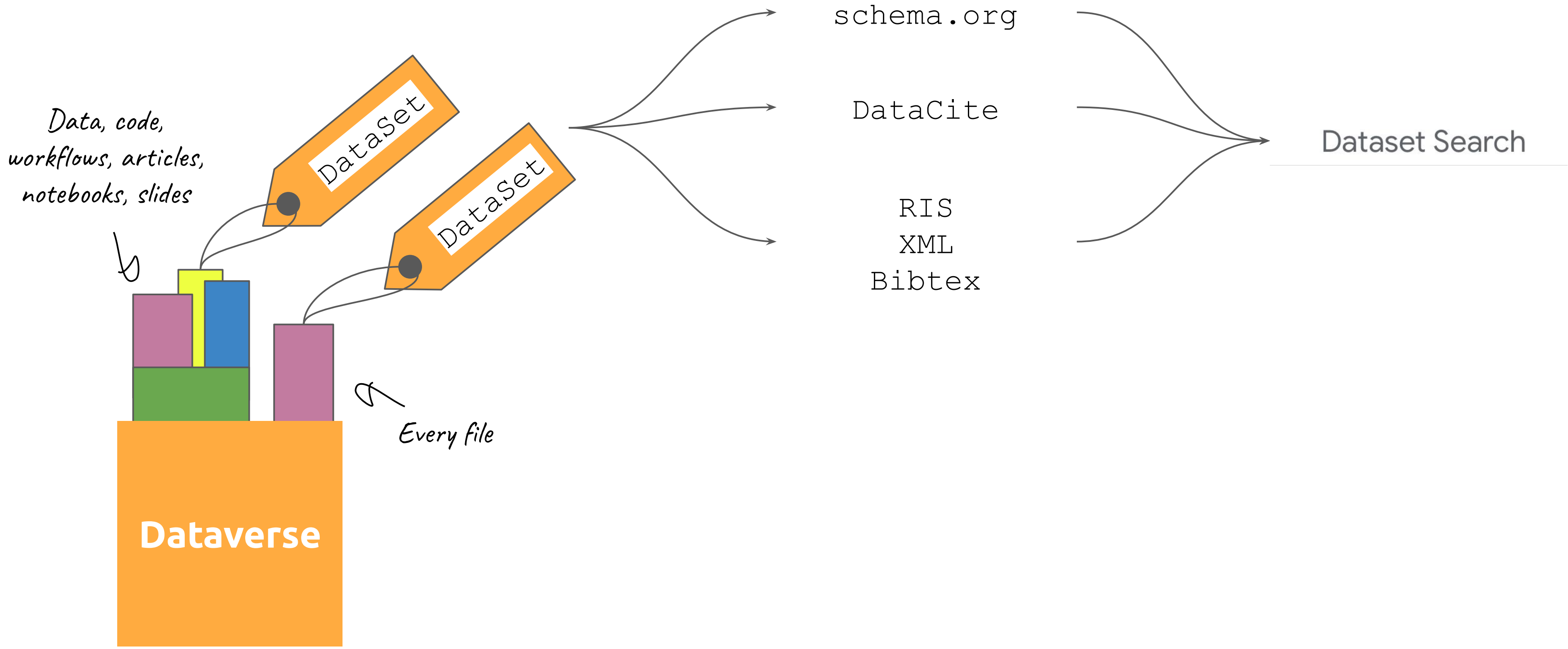
Contact *: Name: Trisovic, Ana — Affiliation: Harvard University
E-mail *: anatrisovic@fas.harvard.edu

Description *: This field supports only certain HTML tags.
Text *:
Date: YYYY-MM-DD

Subject *: Select...

Keyword: Term: — Vocabulary:
Vocabulary URL: Enter full URL, starting with http://

Related Publication: Citation:



Data, code, workflows, articles, notebooks, slides

DataSet

DataSet

Every file

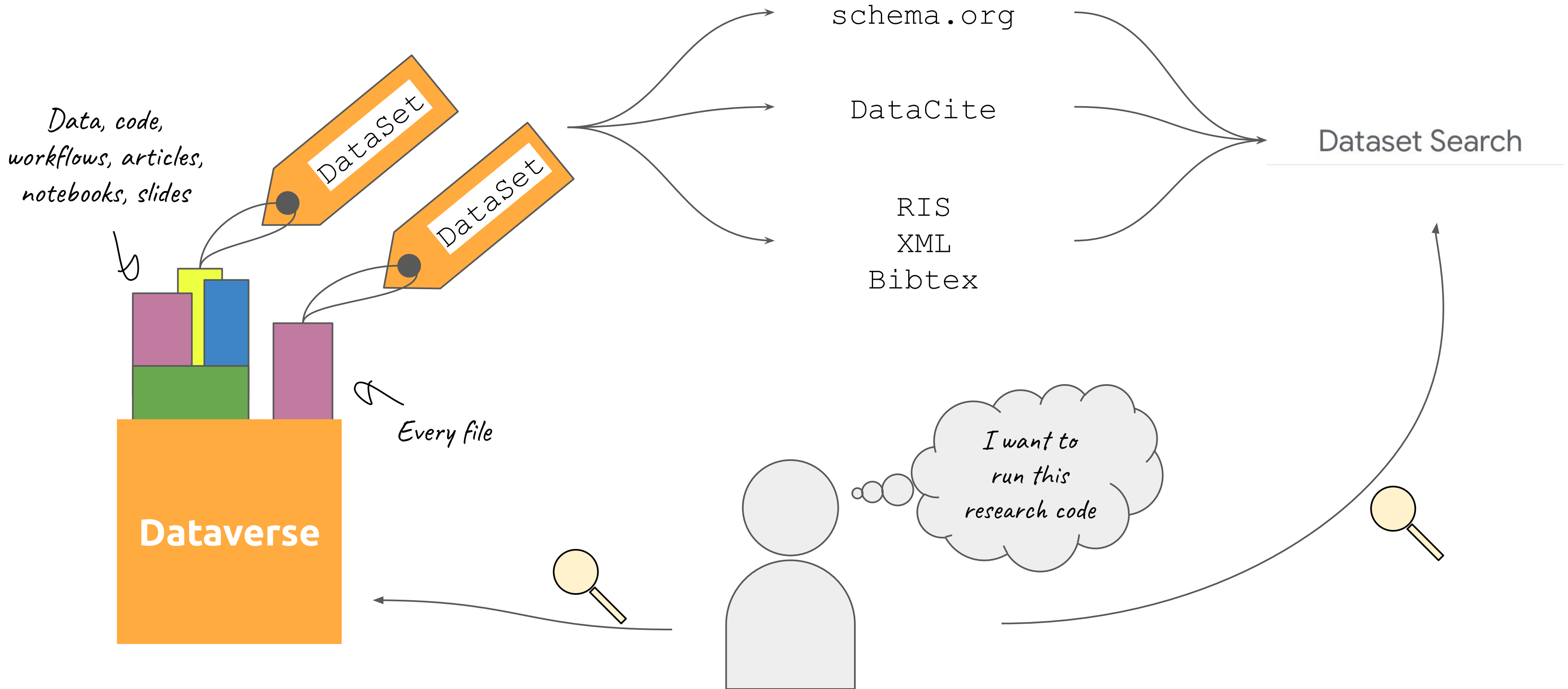
Dataverse

schema.org

DataCite

RIS
XML
Bibtex

Dataset Search



Data, code, workflows, articles, notebooks, slides

DataSet

DataSet

Every file

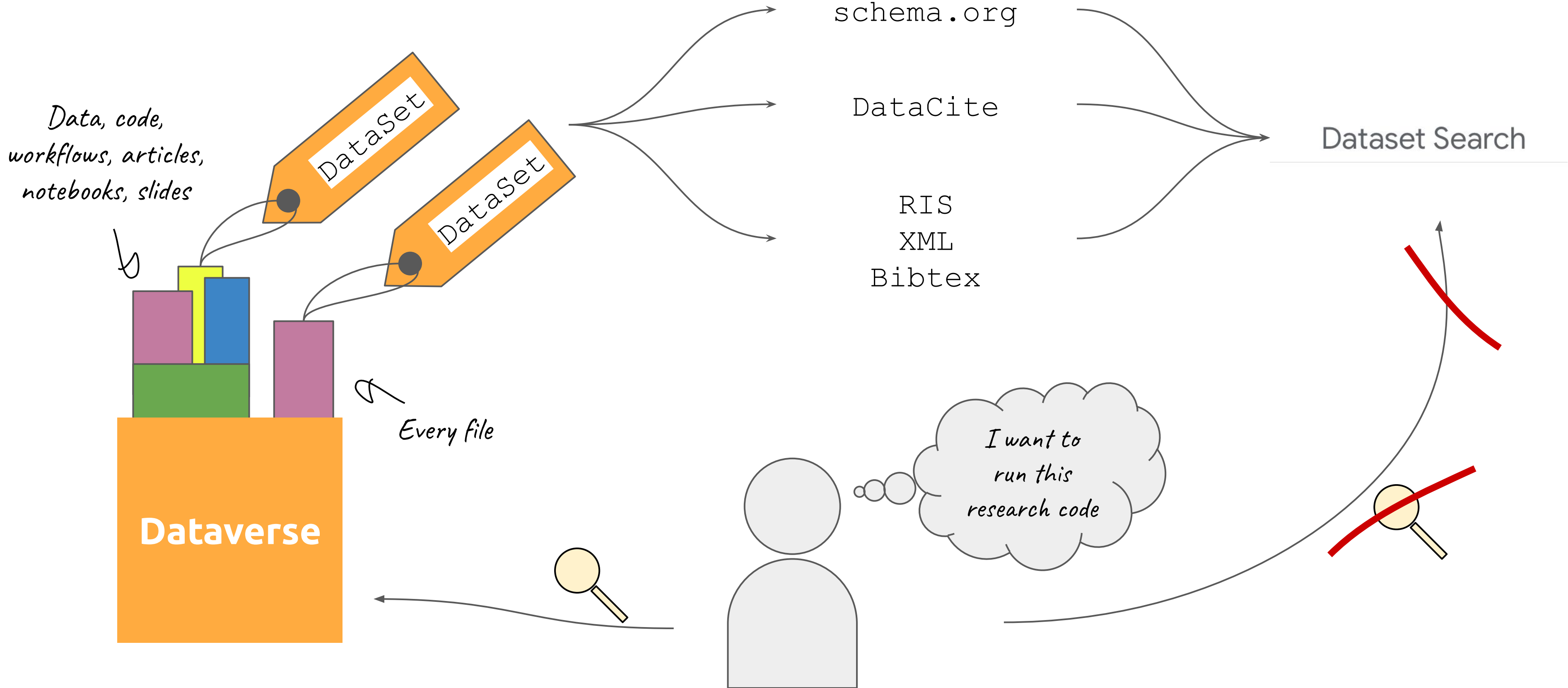
Dataverse

schema.org

DataCite

RIS
XML
Bibtex

Dataset Search



Data, code, workflows, articles, notebooks, slides

DataSet

DataSet

Every file

Dataverse

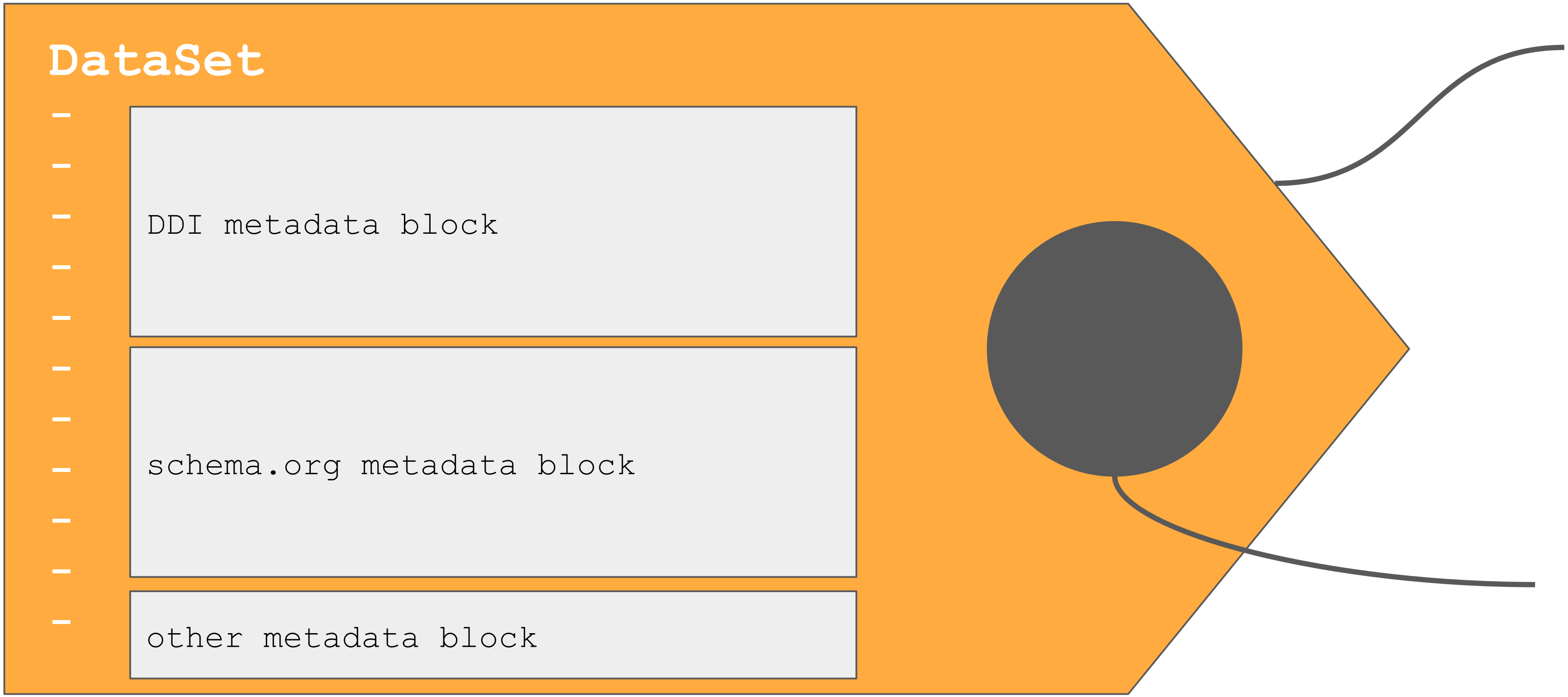
schema.org

DataCite

RIS
XML
Bibtex

Dataset Search

I want to run this research code



DataSet

DDI metadata block

schema.org metadata block

other metadata block

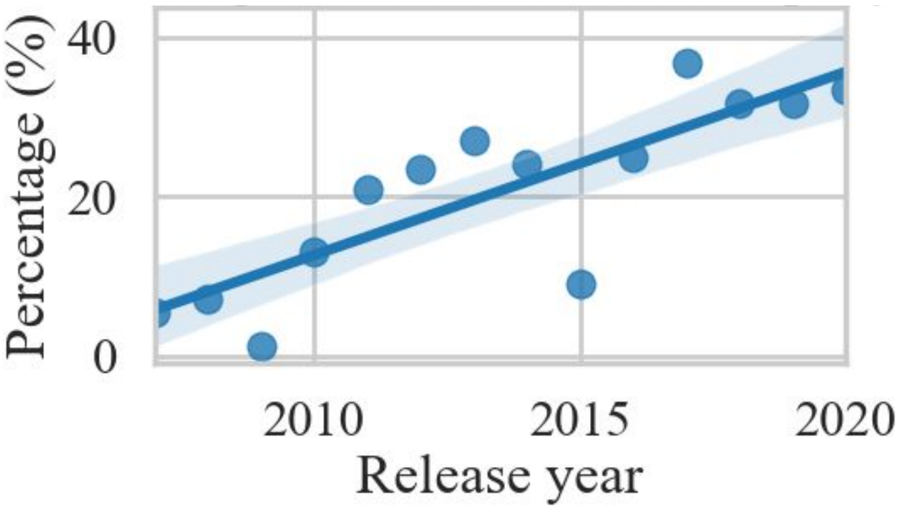
Key problems

1. Data and software licensing
2. Attribution for software contributors
3. Support for new types of files (i.e., container and workflow files)
4. Facilitating software deposit for computational reproducibility
5. Documentation and transparency of research results

Every year we observe higher share of datasets with research code

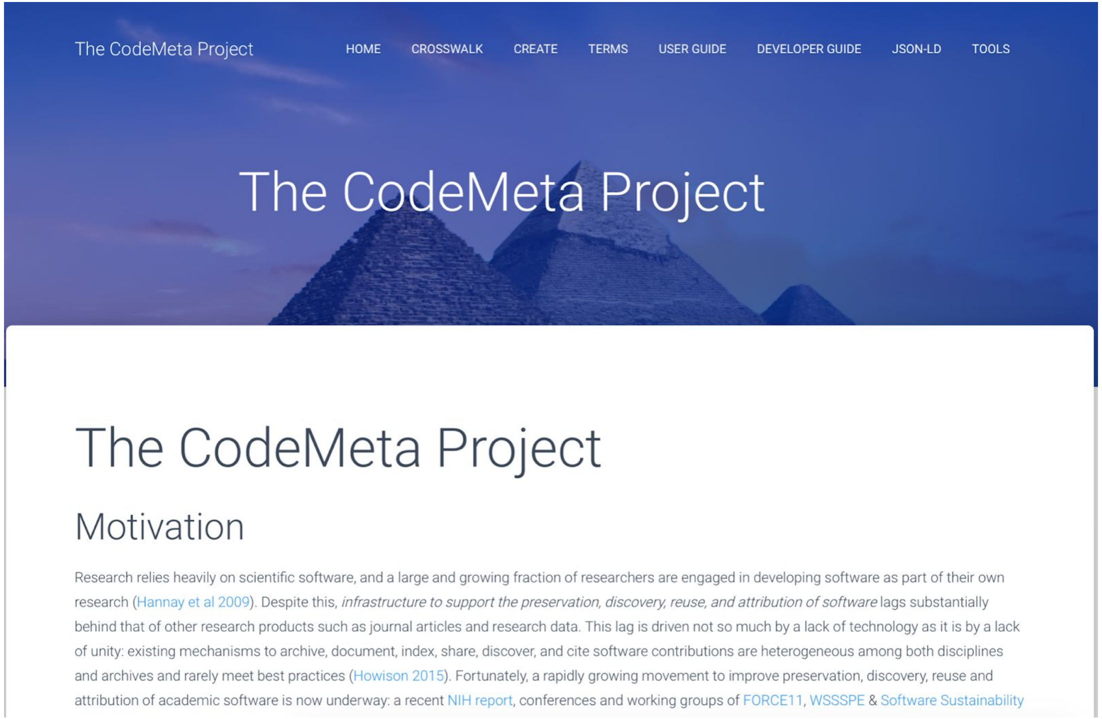
Research code is code files (or scripts) developed by researchers with a goal of obtaining scientific insights

Percentage of datasets with code per year on the Harvard Dataverse repository



Codemeta

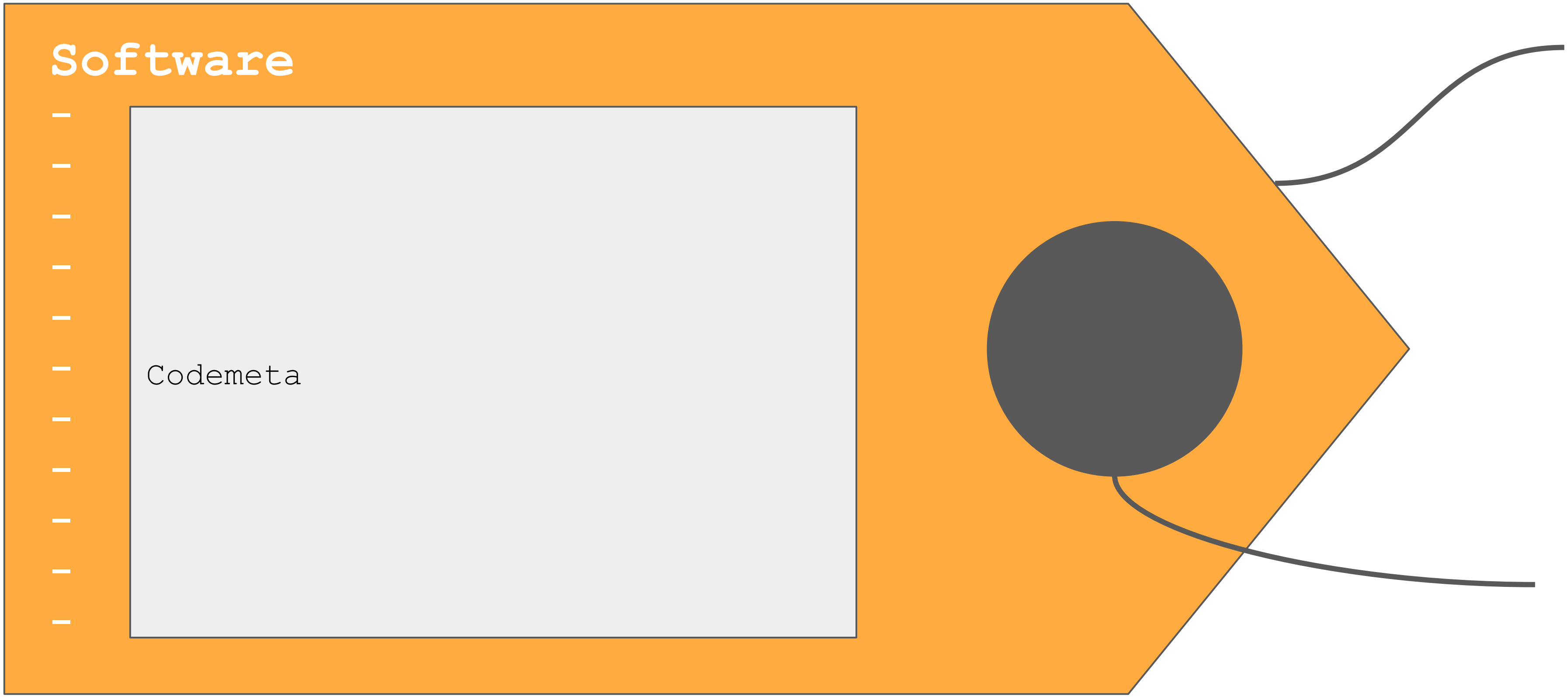
- Specialized metadata schema for research code based on schema.org
- Code attribution for its developers
- Code licensing (different from data licensing)



The CodeMeta Project

Motivation

Research relies heavily on scientific software, and a large and growing fraction of researchers are engaged in developing software as part of their own research (Hannay et al 2009). Despite this, *infrastructure to support the preservation, discovery, reuse, and attribution of software* lags substantially behind that of other research products such as journal articles and research data. This lag is driven not so much by a lack of technology as it is by a lack of unity: existing mechanisms to archive, document, index, share, discover, and cite software contributions are heterogeneous among both disciplines and archives and rarely meet best practices (Howison 2015). Fortunately, a rapidly growing movement to improve preservation, discovery, reuse and attribution of academic software is now underway: a recent NIH report, conferences and working groups of FORCE11, WSSSPE & Software Sustainability



Software

Codemeta

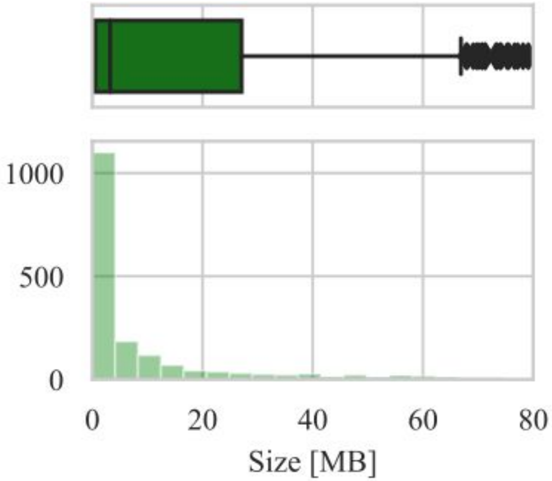
What is the nature of open research data?

Cluster analysis of Dataverse datasets

- The goal is to identify common groups of datasets to inform future developments on metadata
- For instance, we expect to see clusters of:
    - Data & code bundles
    - Code & text files (ie, software)
- Sample size: 45,000 datasets from Harvard Dataverse

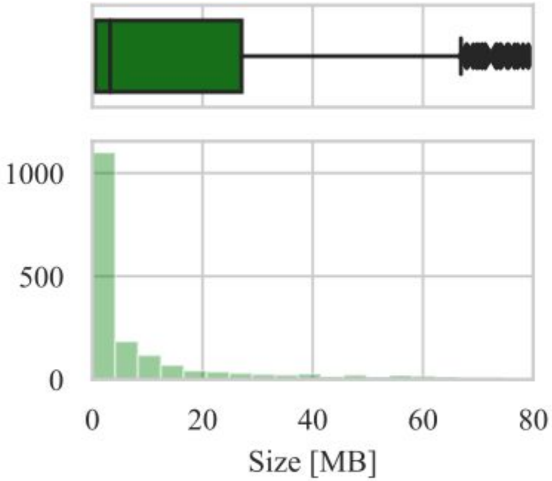
Replication dataset

Distribution of dataset sizes

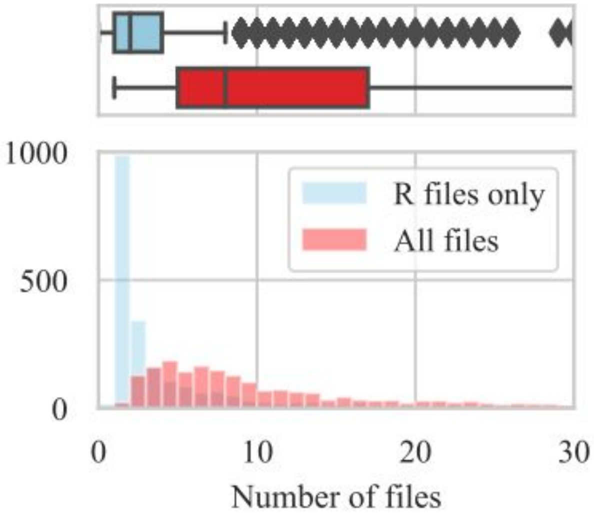


Replication dataset

Distribution of dataset sizes

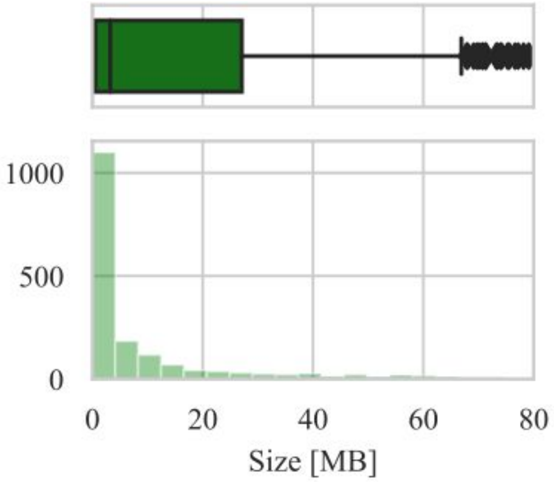


Number of files per dataset

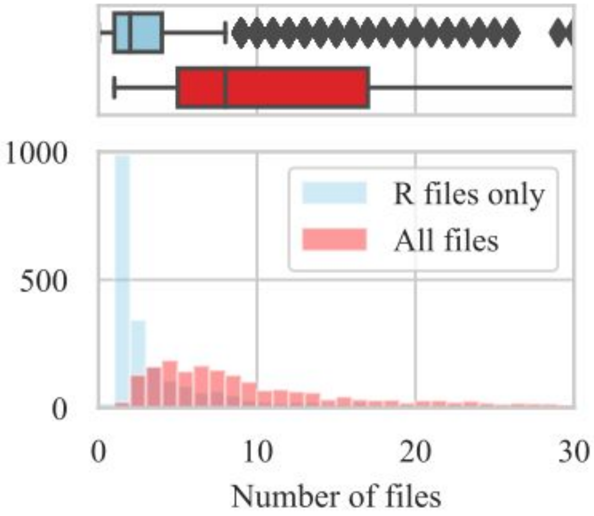


Replication dataset

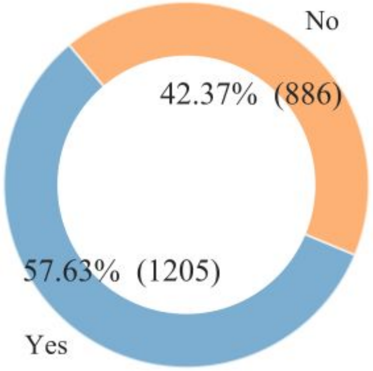
Distribution of dataset sizes



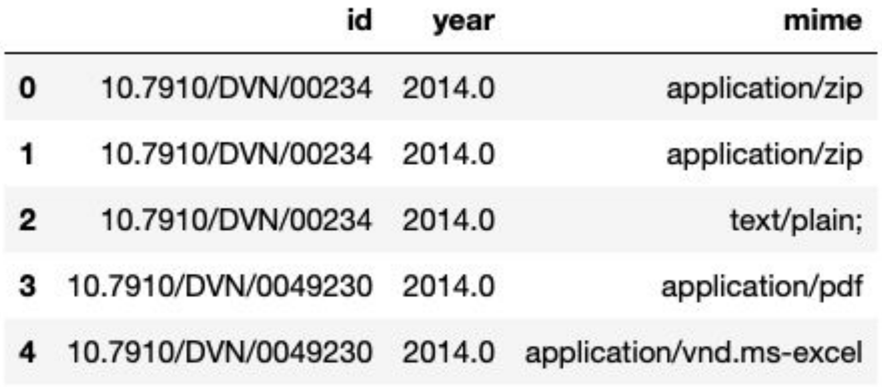
Number of files per dataset



Dataset contains documentation (readme, codebook or instructions)?

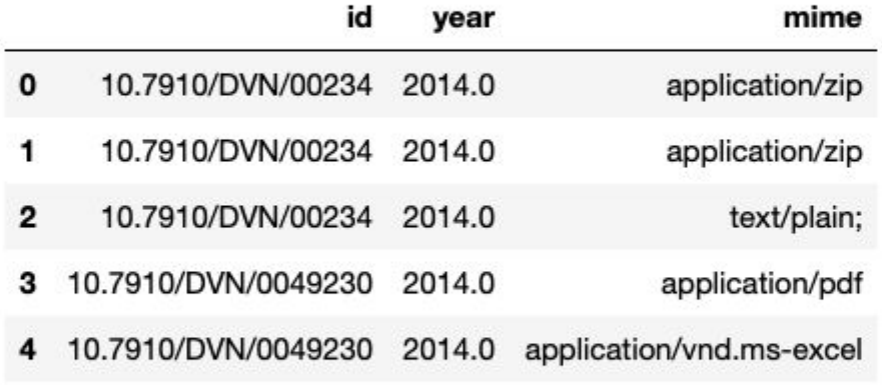


Input data



	id	year	mime
0	10.7910/DVN/00234	2014.0	application/zip
1	10.7910/DVN/00234	2014.0	application/zip
2	10.7910/DVN/00234	2014.0	text/plain;
3	10.7910/DVN/0049230	2014.0	application/pdf
4	10.7910/DVN/0049230	2014.0	application/vnd.ms-excel

Input data

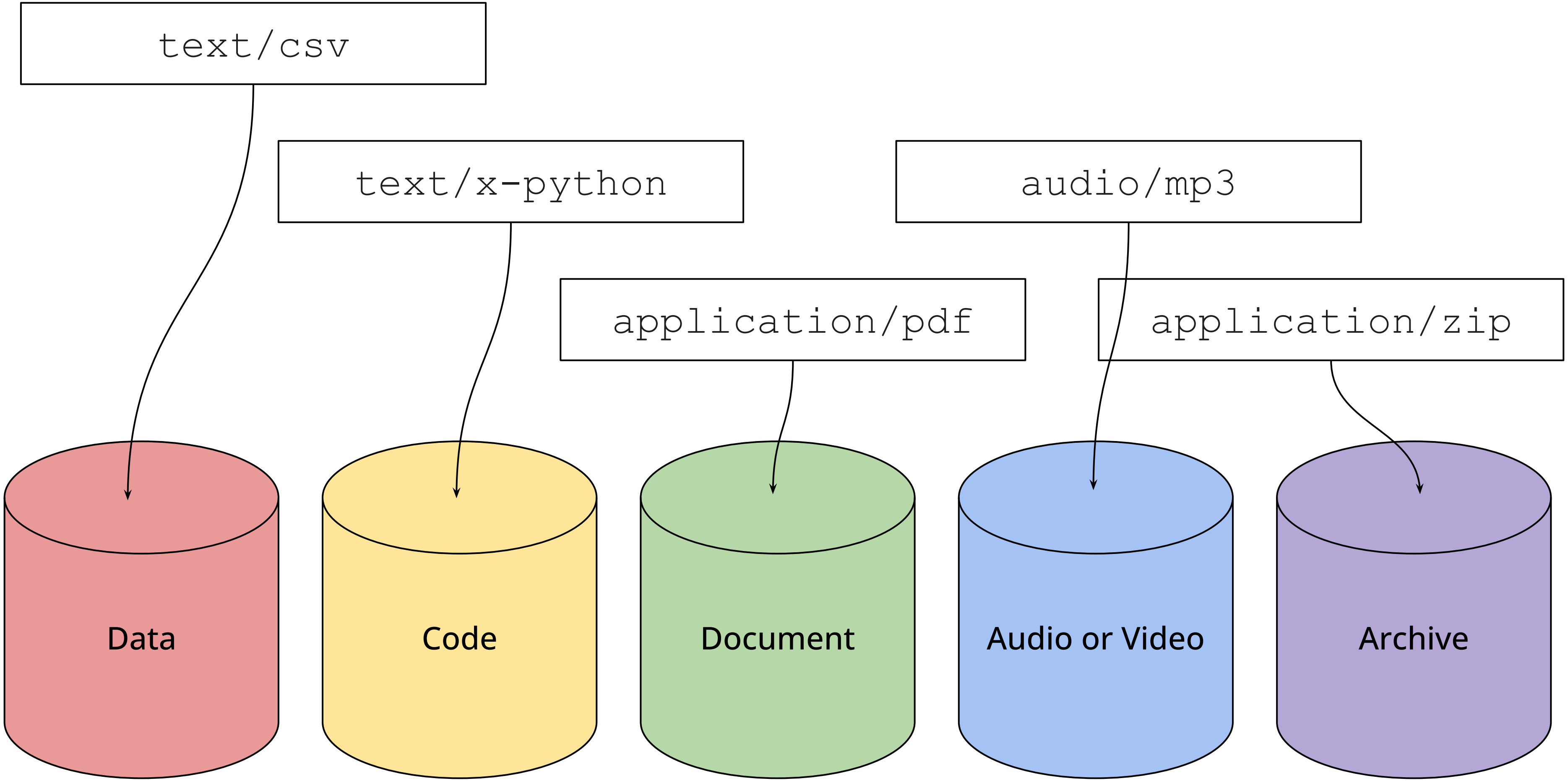


	id	year	mime
0	10.7910/DVN/00234	2014.0	application/zip
1	10.7910/DVN/00234	2014.0	application/zip
2	10.7910/DVN/00234	2014.0	text/plain;
3	10.7910/DVN/0049230	2014.0	application/pdf
4	10.7910/DVN/0049230	2014.0	application/vnd.ms-excel

DOI: 10.7910/DVN/03823

List of files

```
text/csv
text/tsv
text/x-python
text/x-python
application/pdf
```



text/csv

text/x-python

application/pdf

audio/mp3

application/zip

Data

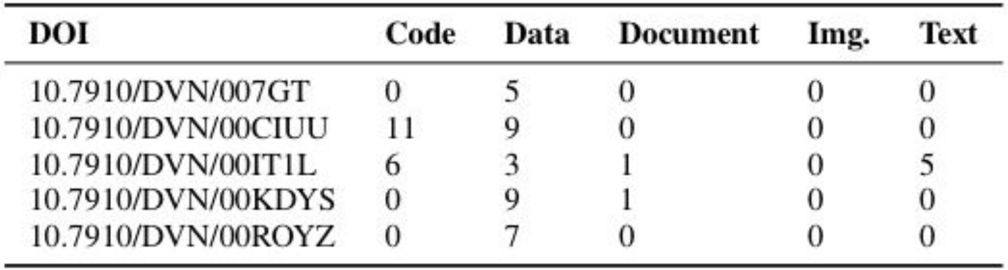
Code

Document

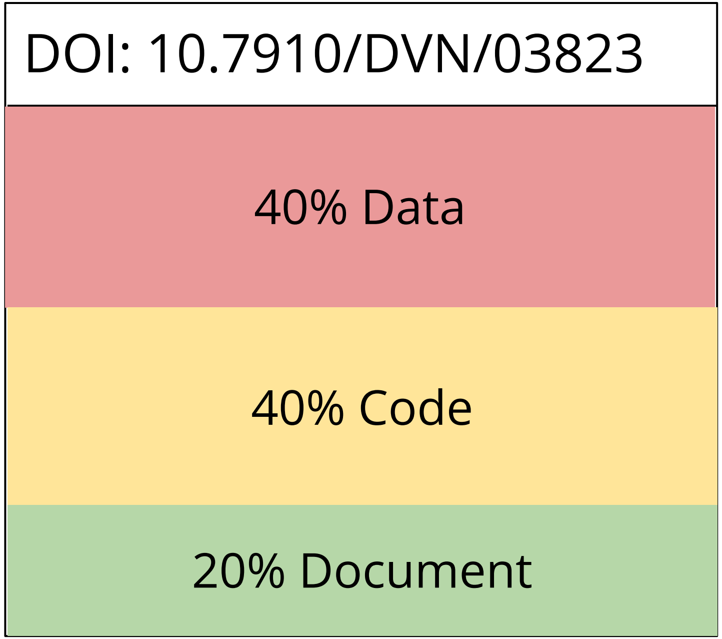
Audio or Video

Archive

Dataset representation



DOI	Code	Data	Document	Img.	Text
10.7910/DVN/007GT	0	5	0	0	0
10.7910/DVN/00CIUU	11	9	0	0	0
10.7910/DVN/00IT1L	6	3	1	0	5
10.7910/DVN/00KDYS	0	9	1	0	0
10.7910/DVN/00ROYZ	0	7	0	0	0



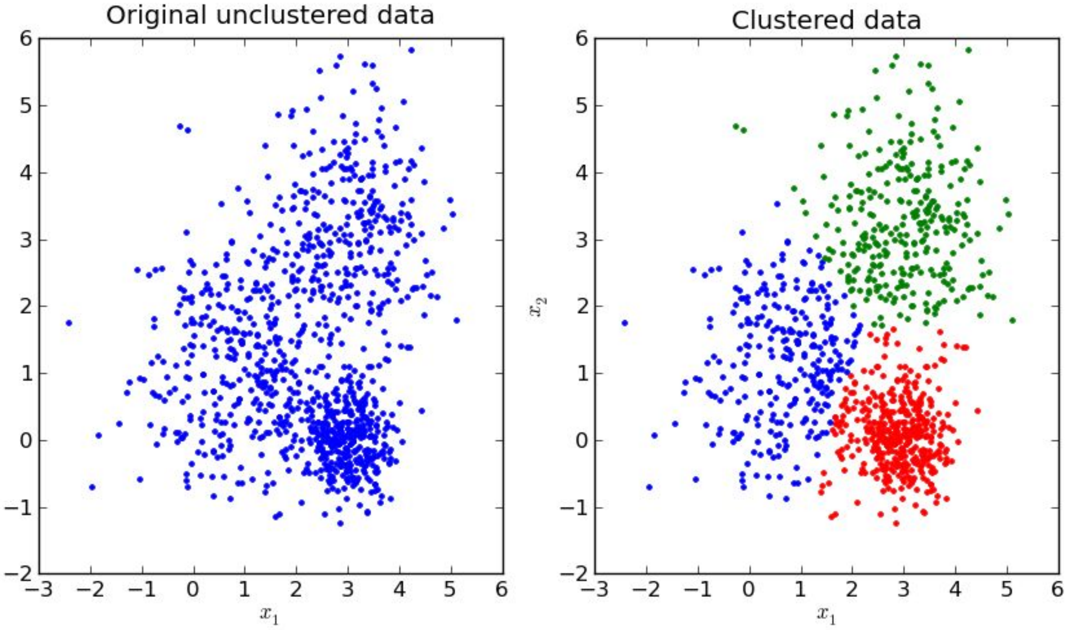
DOI: 10.7910/DVN/03823

40% Data

40% Code

20% Document

- Clustering is a statistical technique where natural grouping within a set are determined, such that the items in each group exhibit more similarly to one another than to items in other groups



Hopkins statistics and the multimodality test

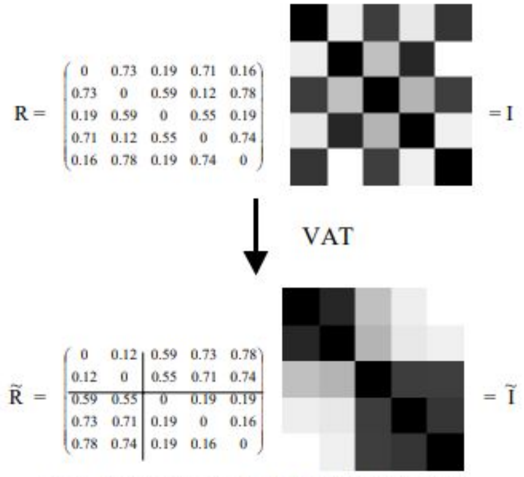
- Application of the clustering algorithms rely on the presence of inherent structure (notion of clusterability).
- The Hopkins statistic is a way of measuring the cluster tendency of a data set.

Hopkins statistics and the multimodality test

- Application of the clustering algorithms rely on the presence of inherent structure (notion of clusterability).
- The Hopkins statistic is a way of measuring the cluster tendency of a data set.
- Multimodality test shows multimodal pairwise distances in the data generated from multiple clusters (otherwise the distribution is unimodal)
- Outcome: our data is highly clusterable!

Visual assessment of (cluster) tendency

- VAT produces an image matrix that can be used for visual assessment of cluster tendency
- It computes dissimilarity matrix and reorders it so that similar objects are close to each other



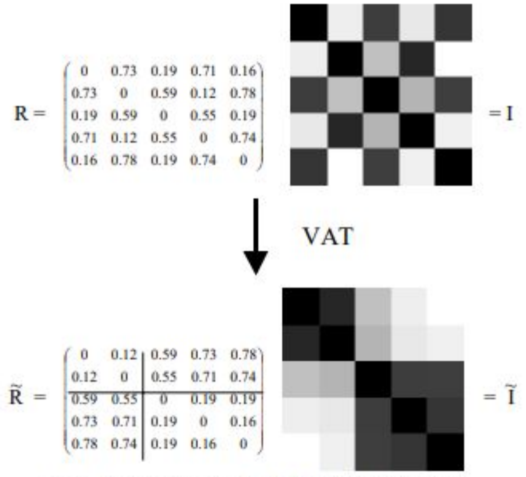
$$\mathrm{R} = \begin{pmatrix} 0 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0 \end{pmatrix} = \mathrm{I}$$

$$\tilde{\mathrm{R}} = \left(\begin{array}{cc|ccc} 0 & 0.12 & 0.59 & 0.73 & 0.78 \\ 0.12 & 0 & 0.55 & 0.71 & 0.74 \\ \hline 0.59 & 0.55 & 0 & 0.19 & 0.19 \\ 0.73 & 0.71 & 0.19 & 0 & 0.16 \\ 0.78 & 0.74 & 0.19 & 0.16 & 0 \end{array}\right) = \tilde{\mathrm{I}}$$

Fig. 5. Results of applying the VAT algorithm to Data Set A.

Visual assessment of (cluster) tendency

- VAT produces an image matrix that can be used for visual assessment of cluster tendency
- It computes dissimilarity matrix and reorders it so that similar objects are close to each other

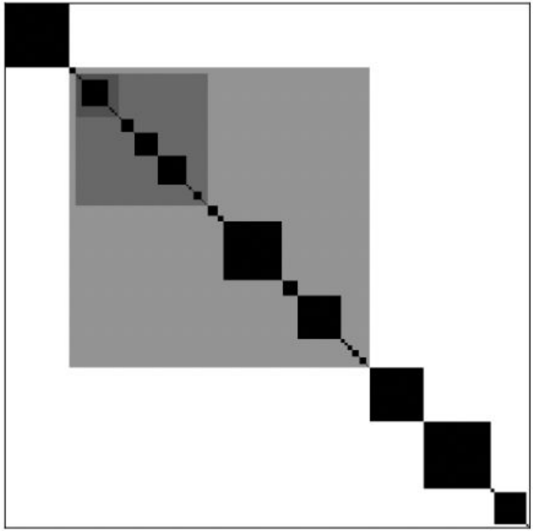


$$\mathrm{R} = \begin{pmatrix} 0 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0 \end{pmatrix}$$

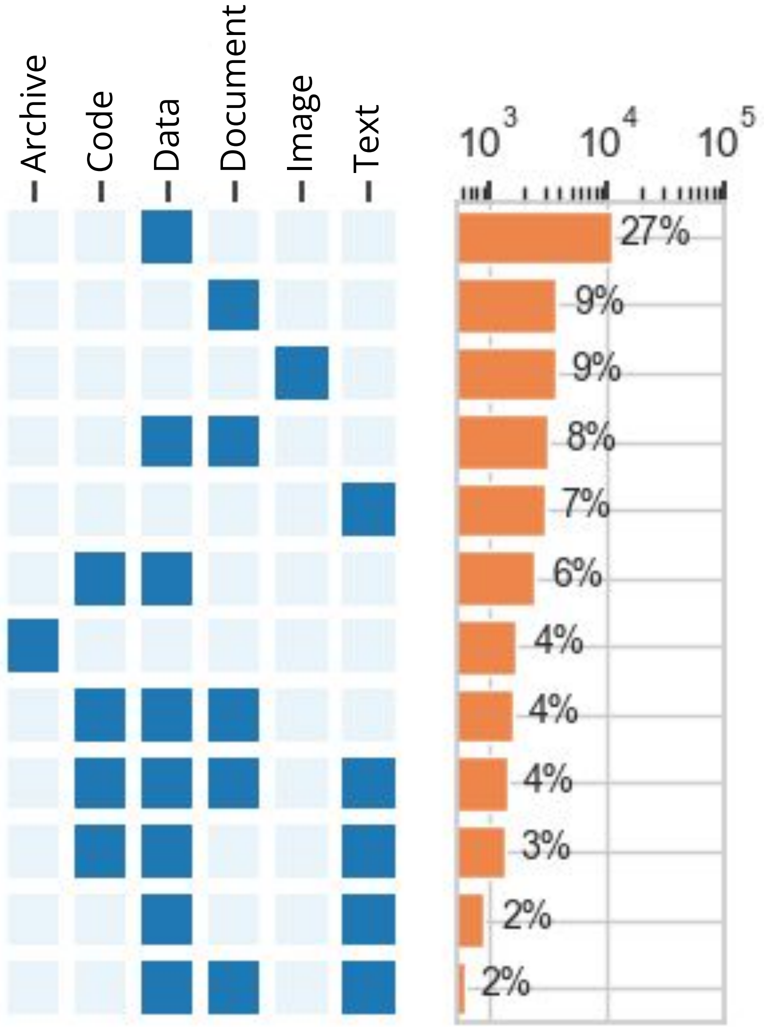
$$\tilde{\mathrm{R}} = \left(\begin{array}{cc|ccc} 0 & 0.12 & 0.59 & 0.73 & 0.78 \\ 0.12 & 0 & 0.55 & 0.71 & 0.74 \\ \hline 0.59 & 0.55 & 0 & 0.19 & 0.19 \\ 0.73 & 0.71 & 0.19 & 0 & 0.16 \\ 0.78 & 0.74 & 0.19 & 0.16 & 0 \end{array}\right)$$

Fig. 5. Results of applying the VAT algorithm to Data Set A.

Outcome:

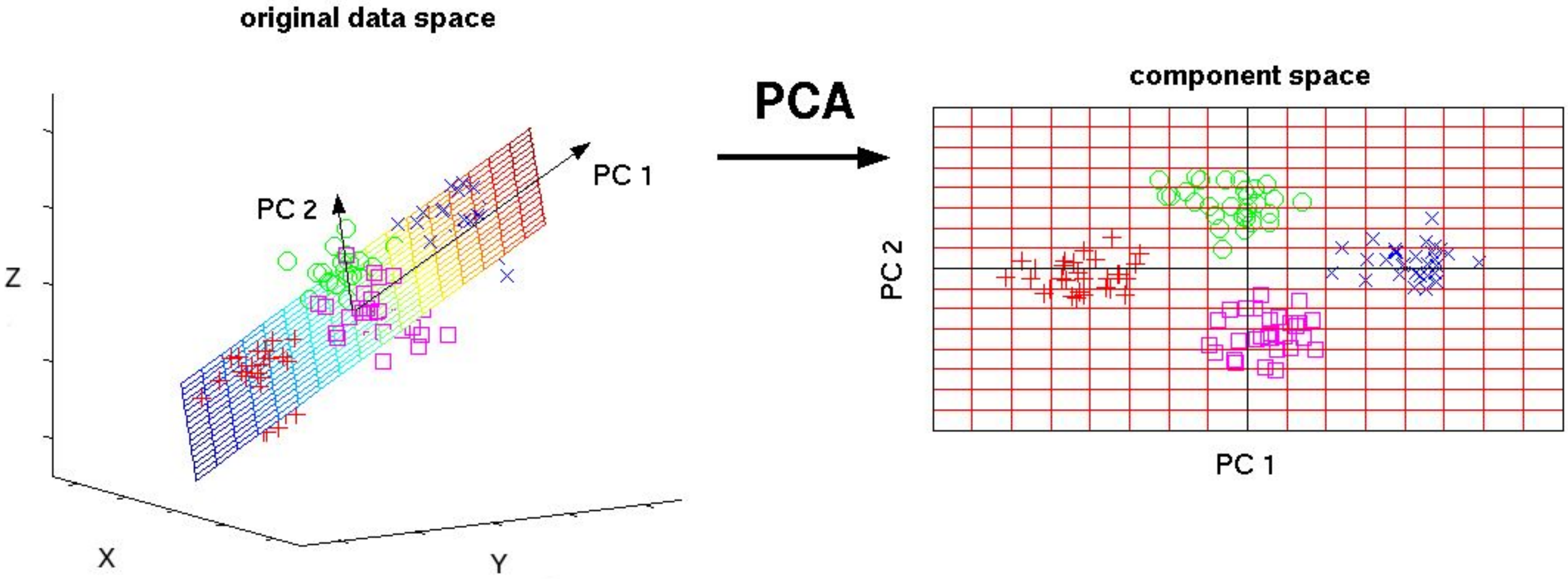


Most frequent dataset content types and their frequencies



Principal component analysis (PCA)

- An unsupervised learning approach used in exploratory analyses that reduces data from high dimensions to lower dimensions while preserving the covariance in the data.

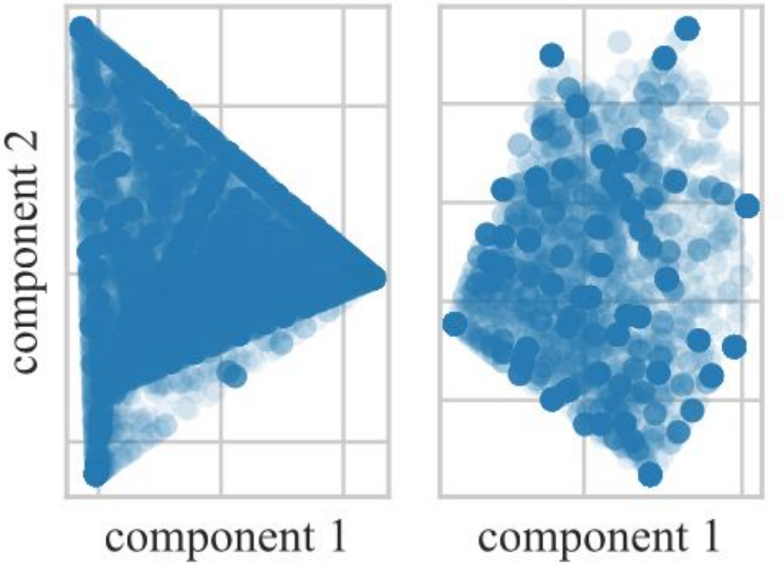

original data space
PC 2
PC 1
Z
X
Y
PCA
component space
PC 2
PC 1

Principal component analysis (PCA)

- An unsupervised learning approach used in exploratory analyses that reduces data from high dimensions to lower dimensions while preserving the covariance in the data.

Principal component analysis (PCA)

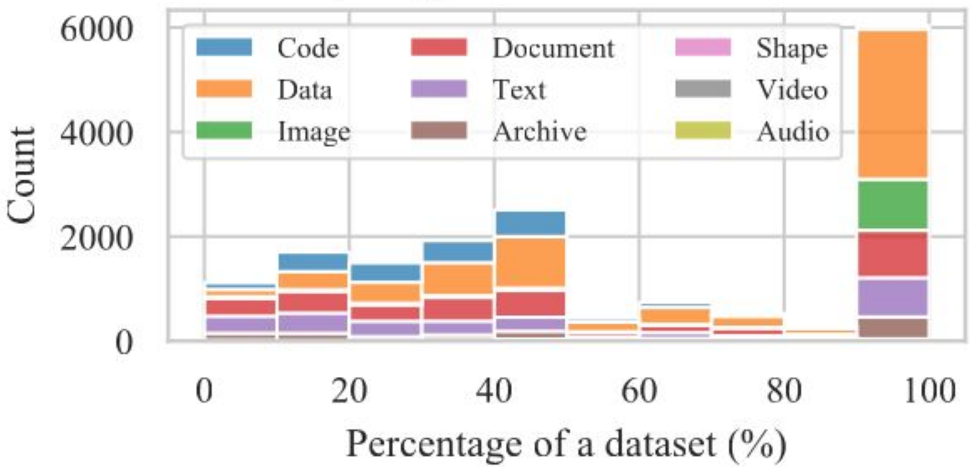
- An unsupervised learning approach used in exploratory analyses that reduces data from high dimensions to lower dimensions while preserving the covariance in the data.

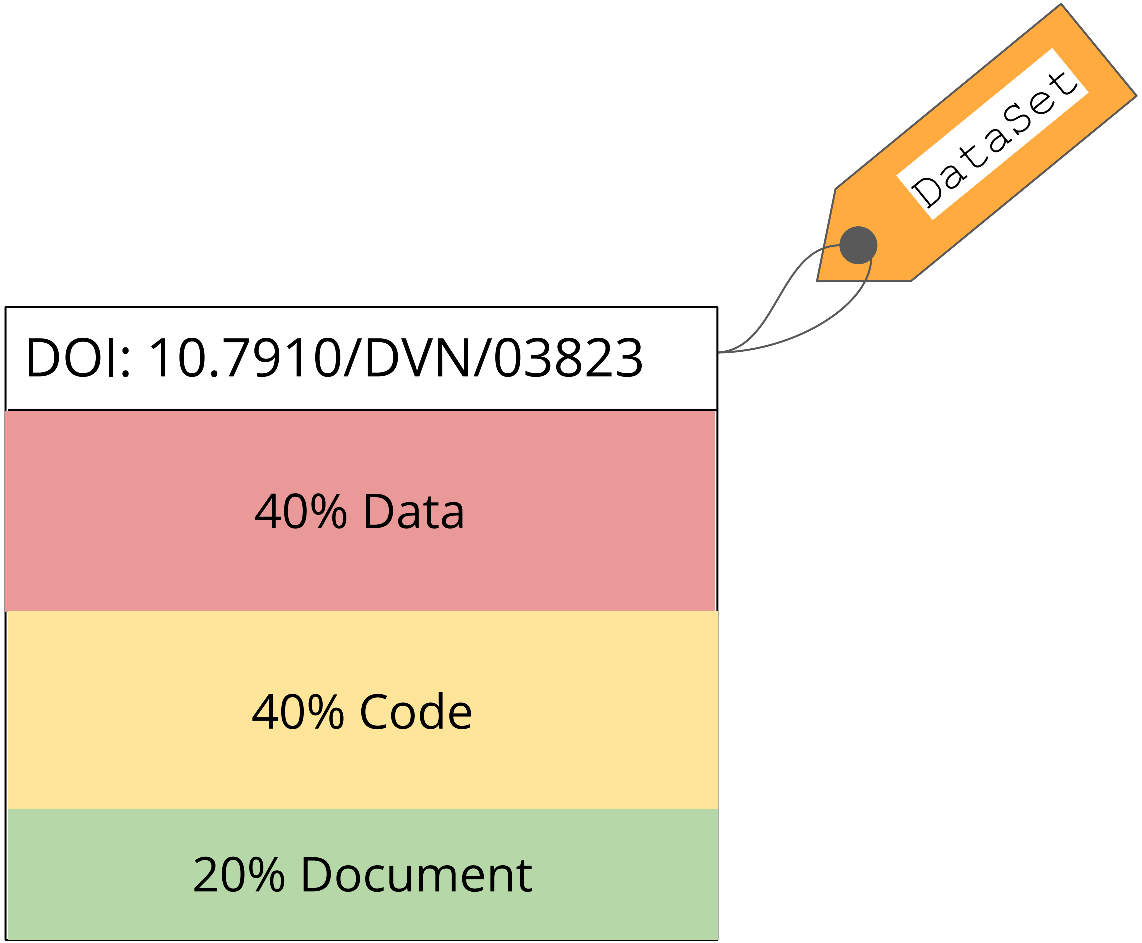


Implications

- Flexibility is necessary when describing existing research datasets (potentially with exception of single-type dataset, i.e., only data, or only code)

Share of object types





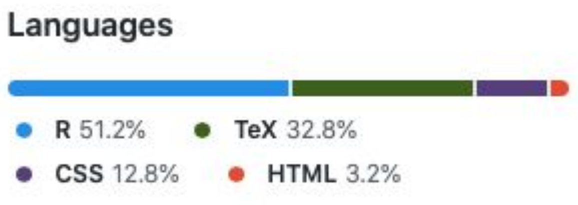
DOI: 10.7910/DVN/03823

40% Data

40% Code

20% Document

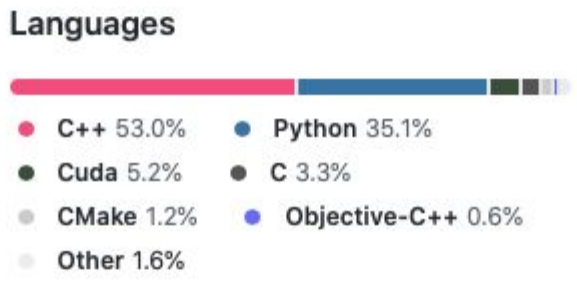
A GitHub solution



Languages

● R 51.2% ● TeX 32.8%

● CSS 12.8% ● HTML 3.2%



Languages

● C++ 53.0% ● Python 35.1%

● Cuda 5.2% ● C 3.3%

● CMake 1.2% ● Objective-C++ 0.6%

● Other 1.6%

RO-Crate as replication metadata

- A bag of references
- Provides an integrated view of resources for FAIR & reproducible research
- Captures unique identifiers, metadata and how they link together (could also include physical objects)
- Each of the packages has its own metadata

📄

ID? Title? Description?

👩‍🔬 Who created this data?

📄 What parts does it have?

📅 When?

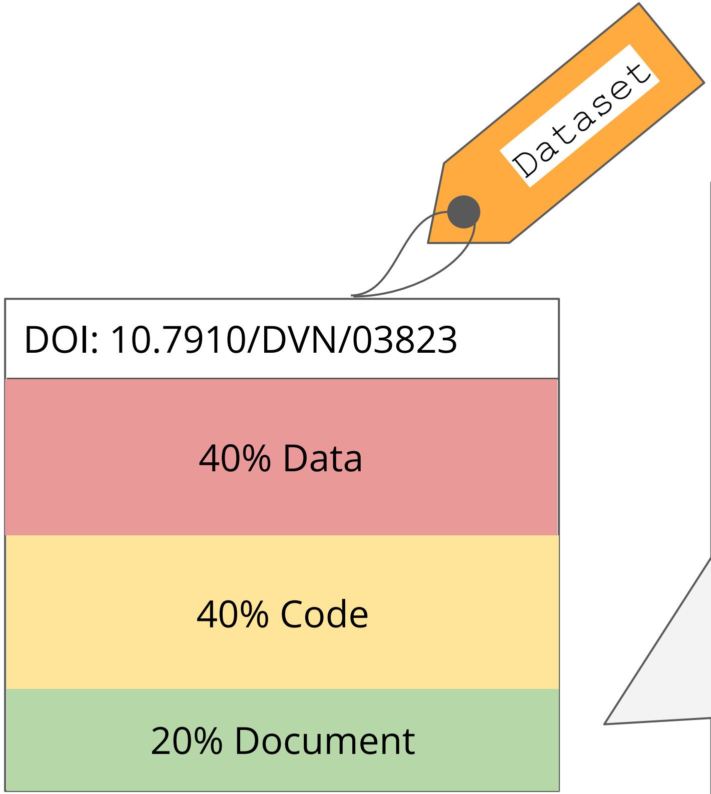
🗒️ What is it about?

♻️ How can it be reused?

🏗️ As part of which project?

💰 Who funded it?

⚒️ How was it made?



Dataset

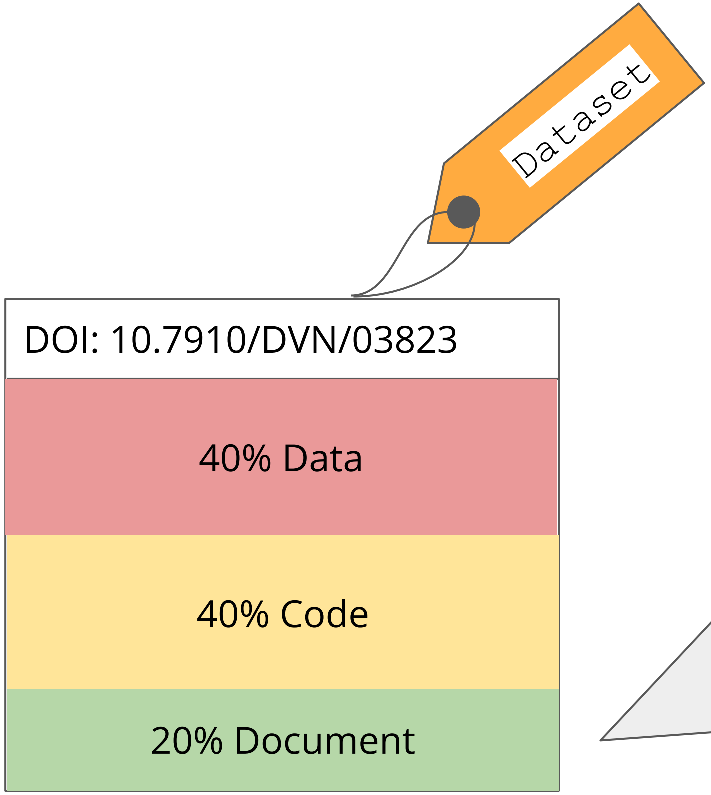
DOI: 10.7910/DVN/03823

40% Data

40% Code

20% Document

```
{
      "@id": "2021-04-08 07.58.17.jpg",
      "@type": "File",
      "contentSize": 3271409,
      "dateModified": "2021-04-08T07:58:17+10:00",
      "description": "",
      "encodingFormat": [
        {
          "@id":
"https://www.nationalarchives.gov.uk/PRONOM/x-fmt/
391"
        },
        "image/jpeg"
      ],
      "name": "Cute puppy"
    },
```

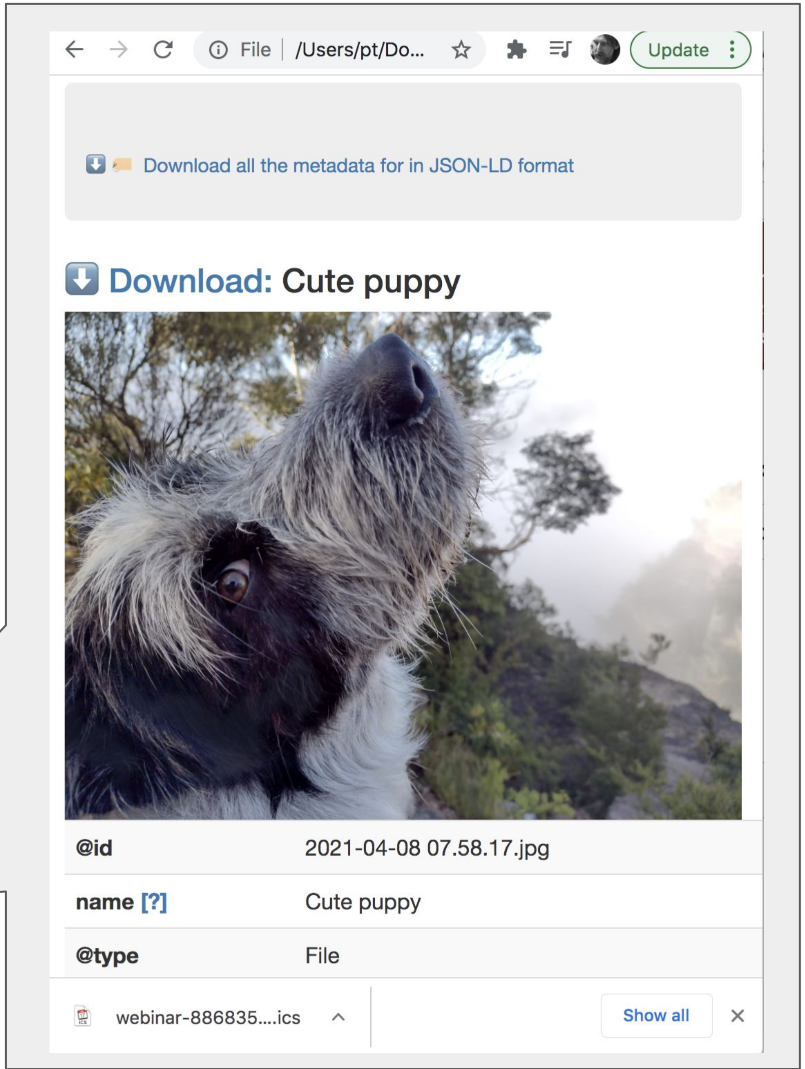


DOI: 10.7910/DVN/03823

40% Data

40% Code

20% Document



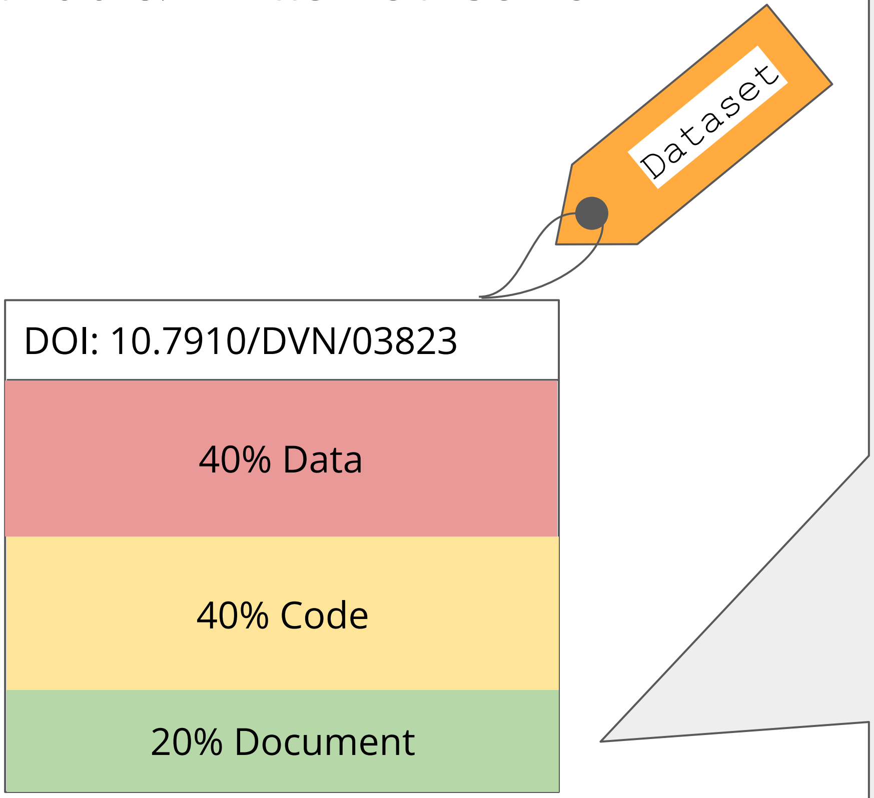
Download all the metadata for in JSON-LD format

Download: Cute puppy

@id	2021-04-08 07.58.17.jpg
name [?]	Cute puppy
@type	File

Email: anatrisovic@g.harvard.edu

GitHub & Twitter: atrisovic



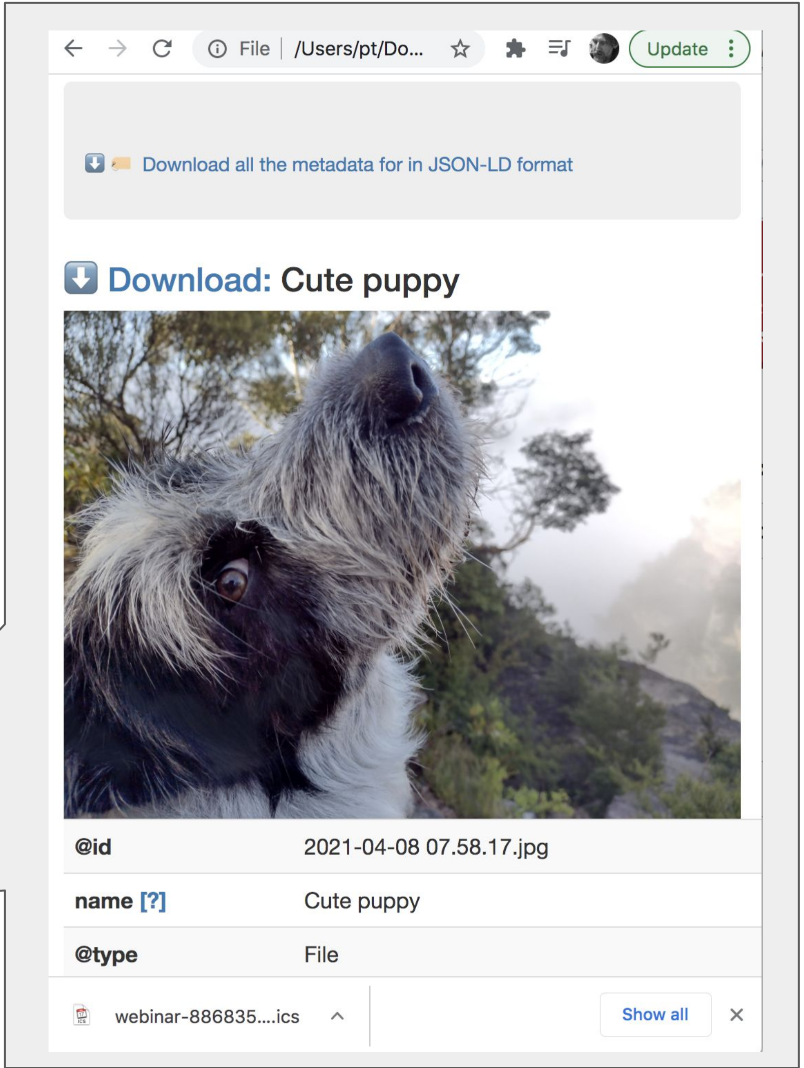
Dataset

DOI: 10.7910/DVN/03823

40% Data

40% Code

20% Document



Download all the metadata for in JSON-LD format

Download: Cute puppy

@id	2021-04-08 07.58.17.jpg
name [?]	Cute puppy
@type	File