

# ITCS 5112: Final Project

## DATASET SEARCH VISUALIZATION

### Group 7

- Saumy Dadhich
- Gaurav Yadav
- Manjunath Shrishail Birajdar
- Hemantha Krishna Chandra Sekhar Chundi

## Requirements:

- ▶ Filtering Dataset on the basis of the number of entries or number of attributes.
- ▶ **Display Dataset Integrity.**
- ▶ **Displaying Popular Categories.**
- ▶ Display Dataset Search Activities.

## Difficulties while extracting the data

1. Eliminating the blank json files during json parsing.
2. Initially the files had only one subject(Social Science).
3. Data required cleaning after parsing and writing to csv.

## Pre Processing

- ▶ Understanding JSON.
- ▶ Identifying relevant keys.
- ▶ Python Script to parse JSON and write to csv.
- ▶ Cleaning the csv file

100



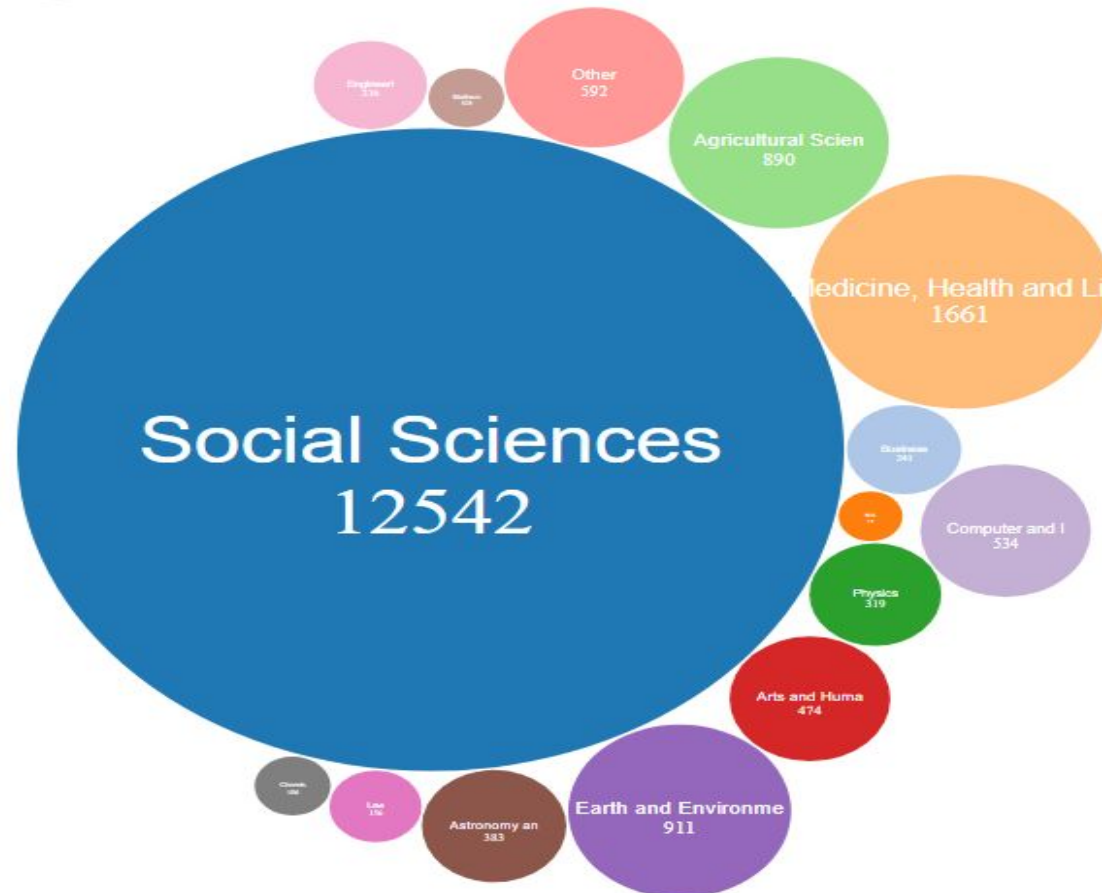
```
1  {"status":"OK",
2    "data":{"id":3316971,
3      "identifier":"DVN/7DTR3E",
4      "persistentUrl":"https://doi.org/10.7910/DVN/7DTR3E",
5      "protocol":"doi",
6      "authority":"10.7910",
7      "publisher":"Harvard Dataverse","publicationDate":"2018-11-18","storageIdentifier":"s3://10.7910/DVN/7DTR3E",
8      "latestVersion":{"id":147279,"storageIdentifier":"s3://10.7910/DVN/7DTR3E","versionNumber":1,"versionMinorNumber":0,"versionState":"RELEASED",
9        "productionDate":"Production Date","lastUpdateTime":"2018-11-19T00:28:43Z","releaseTime":"2018-11-19T00:28:43Z","createTime":"2018-11-19T00:28:37Z","license":"CC0",
10       "termsOfUse":"CC0 Waiver","metadataBlocks":{"citation":{"displayName":"Citation Metadata",
11         "fields":[{"typeName":"title","multiple":false,"typeClass":"primitive","value":"demo-data"},
12           {"typeName":"author","multiple":true,"typeClass":"compound","value":[{"authorName":{"typeName":"authorName","multiple":false,"typeClass":"primitive","value":"Quan, Josh"}
13             • Josh},"datasetContactEmail":{"typeName":"datasetContactEmail","multiple":false,"typeClass":"primitive","value":"joshua.quan@berkeley.edu"}]}]},
14           {"typeName":"dsDescription","multiple":true,"typeClass":"compound","value":[{"dsDescriptionValue":{"typeName":"dsDescriptionValue","multiple":false,"typeClass":"primitive"
15             {"typeName":"dateOfDeposit","multiple":false,"typeClass":"primitive","value":"2018-11-18"}]}]},"files":[]}}}
```

# CSV file

	A	B	C	D	E	F	G	H	I	J
1	title	URL	author	desc	subject	keywords	publication			
2	United Sta	https://do	['Inter-univ	This	[]	['candidates', 'congressional elections', 'counties', 'elec				
3	Candidate	https://do	['Inter-univ	This data	[]	['candidates', 'constituencies', 'elections', 'political attit				
4	Historical,	https://do	['Inter-univ	Detailed co	[]	['census data', 'counties', 'demographic characteristics'				
5	United Sta	https://do	['Inter-univ	Roll call vo	[]	['eighteenth century', 'historical data', 'legislators', 'nine				
6	Data Conf	https://do	['Inter-univ	This study	[]	['congressional elections', 'demographic characteristics				
7	Referenda	https://do	['Inter-univ	This data c	[]	['congressional elections (US House)', 'congressional ele				
8	Federal De	https://do	['Federal C	This data c	[]	['bank deposits', 'banks', 'counties', 'economic history',				
9	Censuses c	https://do	['United St	This data	[]	['census data', 'church membership', 'counties', 'religion				
10	Farm Real	https://do	['Pressley,	This data c	[]	['counties', 'farms', 'nineteenth century', 'property value				
11	United Sta	https://do	['United St	This study	[]	['census data', 'congressional districts', 'congressional e				
12	United Sta	https://do	['United St	This study	[]	['candidates', 'congressional districts', 'congressional el				
13	County an	https://do	['United St	Several da	[]	['census data', 'cities', 'counties', 'demographic charact				
14	General El	https://do	['Inter-univ	This data c	[]	['congressional elections', 'election returns', 'elections',				
15	Survey of	https://do	['National	This data c	[]	['church membership', 'counties', 'religion', 'religious co				



# Category Distribution



Packed  
Bubbles

# Redirection on click

**HARVARD**  
Dataverse

Search ▾ About User Guide Support Sign Up Log In

Metrics 6,918,755 Downloads

Contact Share

Search this dataverse... Find Advanced Search + Add Data

☒ Datasets (1,781)  
☒ Datasets (42,132)  
☐ Files (38)

**Dataverse Category**  
Research Project (651)  
Researcher (604)  
Organization or Institution (136)  
Research Group (123)  
Journal (57)  
[More...](#)

**Metadata Source**  
Harvested (24,547)  
Harvard Dataverse (19,366)

**Publication Year**

**Subject: Agricultural Sciences** ✕

1 to 10 of 43,913 Results

Appendix for Marina E. Henke, "Buying Allies: Payment Practices in Multilateral Military Vol. 43, No. 4 (Spring 2019), pp. 128-162, doi.org/10.1162/ISEC\_a\_00345  
Apr 30, 2019 - International Security Dataverse

Henke, Marina, 2019, "Appendix for Marina E. Henke, "Buying Allies: Payment Practices in Multilateral Military Coalition-Building," International Security, Vol. 43, No. 4 (Spring 2019), pp. 128-162, doi.org/10.1162/ISEC\_a\_00345", https://doi.org/10.7910/DVN/TDWN09, Harvard Dataverse, V1

Appendix for Marina E. Henke, "Buying Allies: Payment Practices in Multilateral Military Coalition-Building," International Security, Vol. 43, No. 4 (Spring 2019), pp. 128-162, doi.org/10.1162/ISEC\_a\_00345. This appendix contains supplementary information and the "Pivotal State C...

Soil profile pit characterization  
Apr 30, 2019 - Long-Term Agroforestry Trial - Kenya Dataverse

Njoroge, Julius; Muthuri, Catherine, 2019, "Soil profile pit characterization", https://doi.org/10.7910/DVN/ZYW99C, Harvard Dataverse, V1

Search Results based on Individual Category



# Data Integrity Check

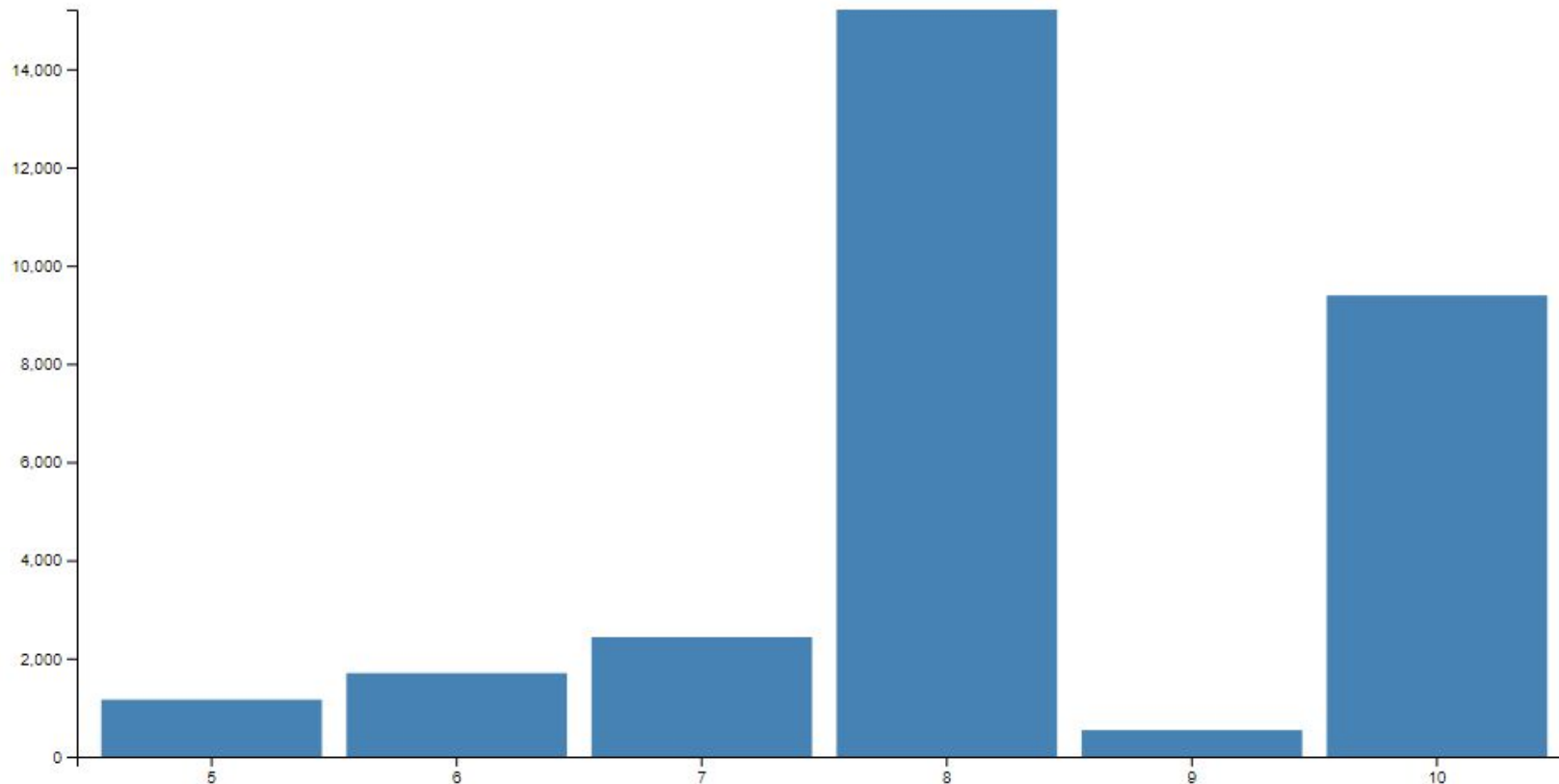
Attributes Considered:

1. Author
2. Description
3. Subject
4. Keywords
5. Publication

Maximum score of 2  
per attribute

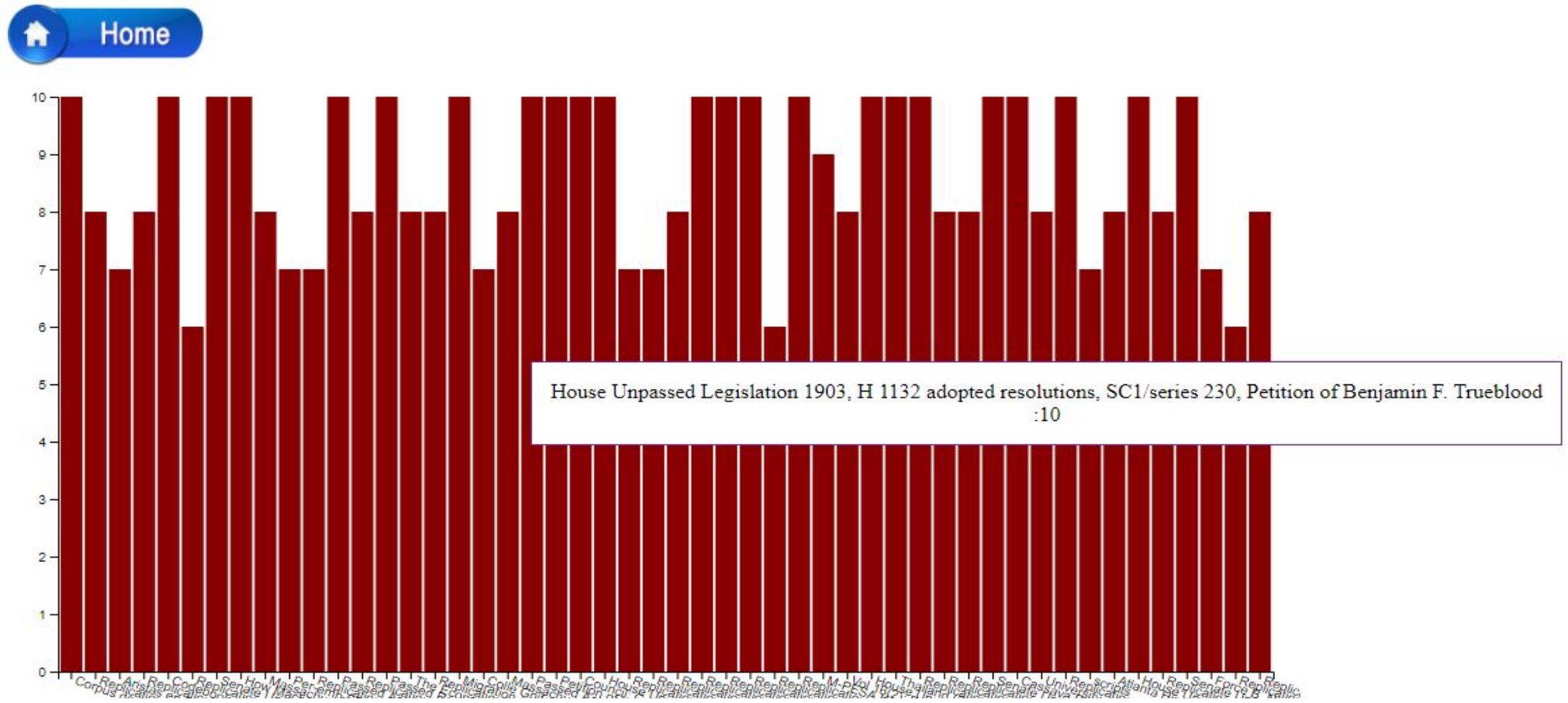
**Total maximum score for each  
dataset is 10.**

# Dataset Count Based on Data Integrity Score



Sided Bar  
Charts

100



Demo

# Limitations

- Time lag while loading visualizations.
- Filters in-between the visualization.
- Make Integrity score more readily available in other visualization.
- Dashboard.



## Conclusion

**learned from our own experience with dataverse**

- ▶ **Categories and linking them directly.**
- ▶ **Use Integrity score to eliminate unwanted datasets.**



THANK YOU !