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

Understanding JSON

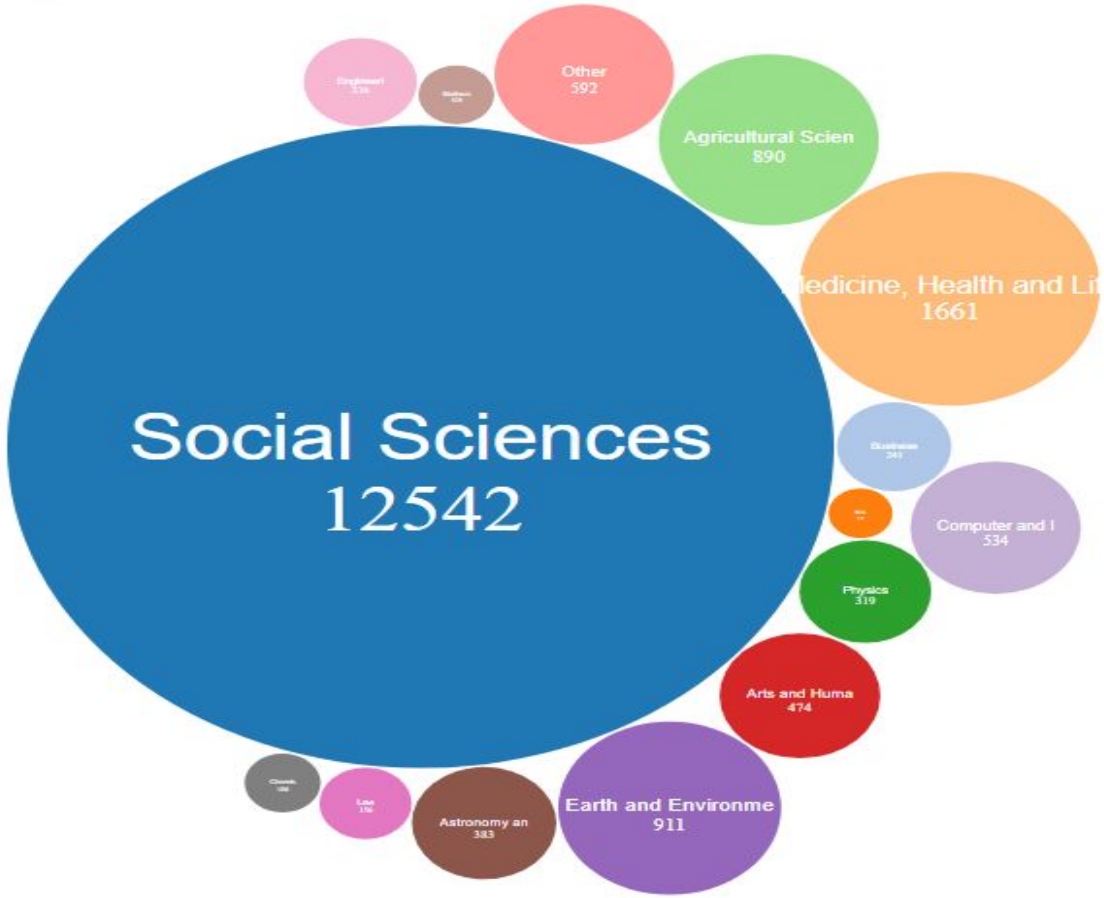
10.7910_DVN_7DTR3E.json


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

CSV file

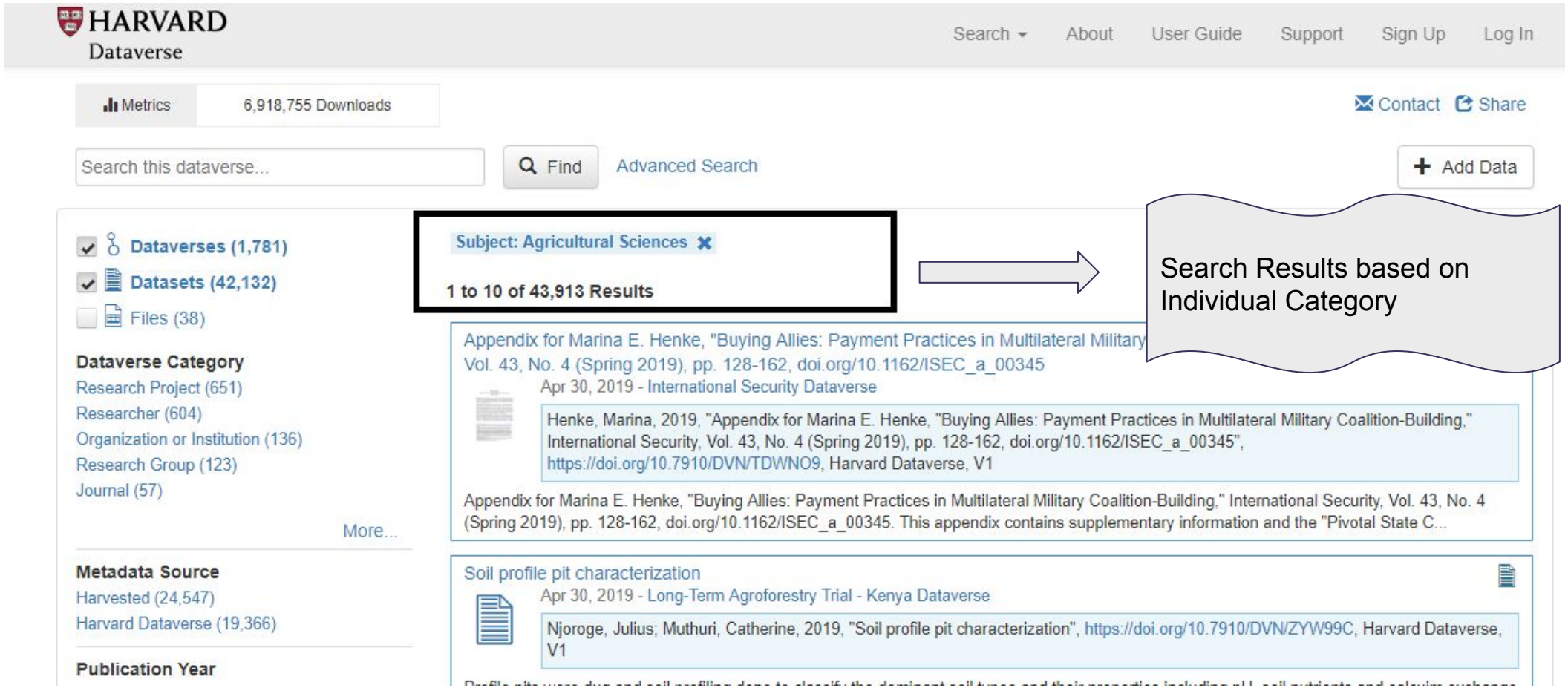
	A	B	C	D	E	F	G	H	I	J
1	title	URL	author	desc	subject	keywords	publication			
2	United Sta	https://do	['Inter-univ	This	[]	['candidates', 'congressional elections', 'counties', 'elec				
3	Candidate	https://do	['Inter-univ	This data	[]	['candidates', 'constituencies', 'elections', 'political attit				
4	Historical,	https://do	['Inter-univ	Detailed c	[]	['census data', 'counties', 'demographic characteristics'				
5	United Sta	https://do	['Inter-univ	Roll call vc	[]	['eighteenth century', 'historical data', 'legislators', 'nine				
6	Data Confi	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 D	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



Data Integrity Check

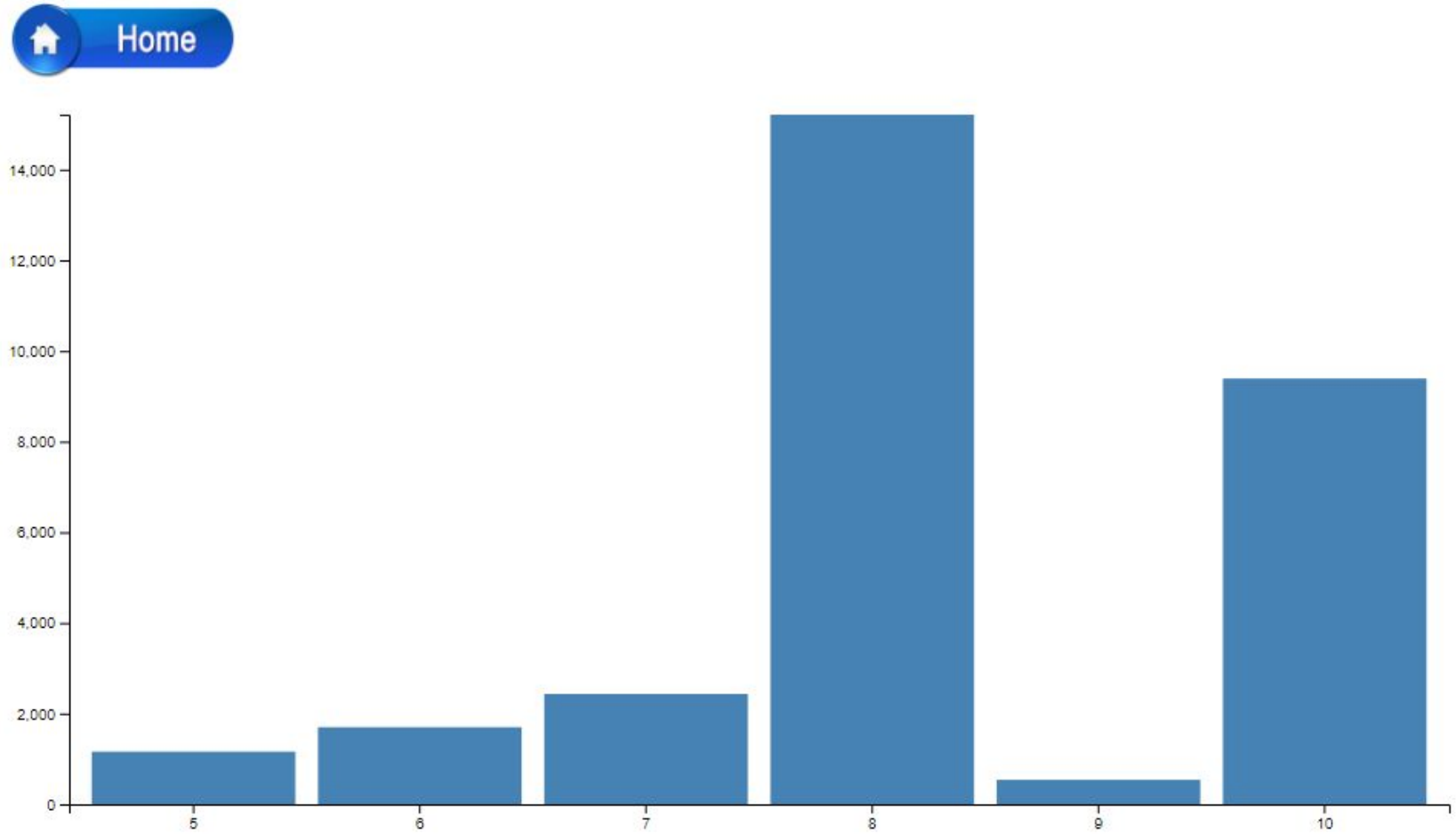
Attributes Considered:

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

Maximum score of 2 per attribute

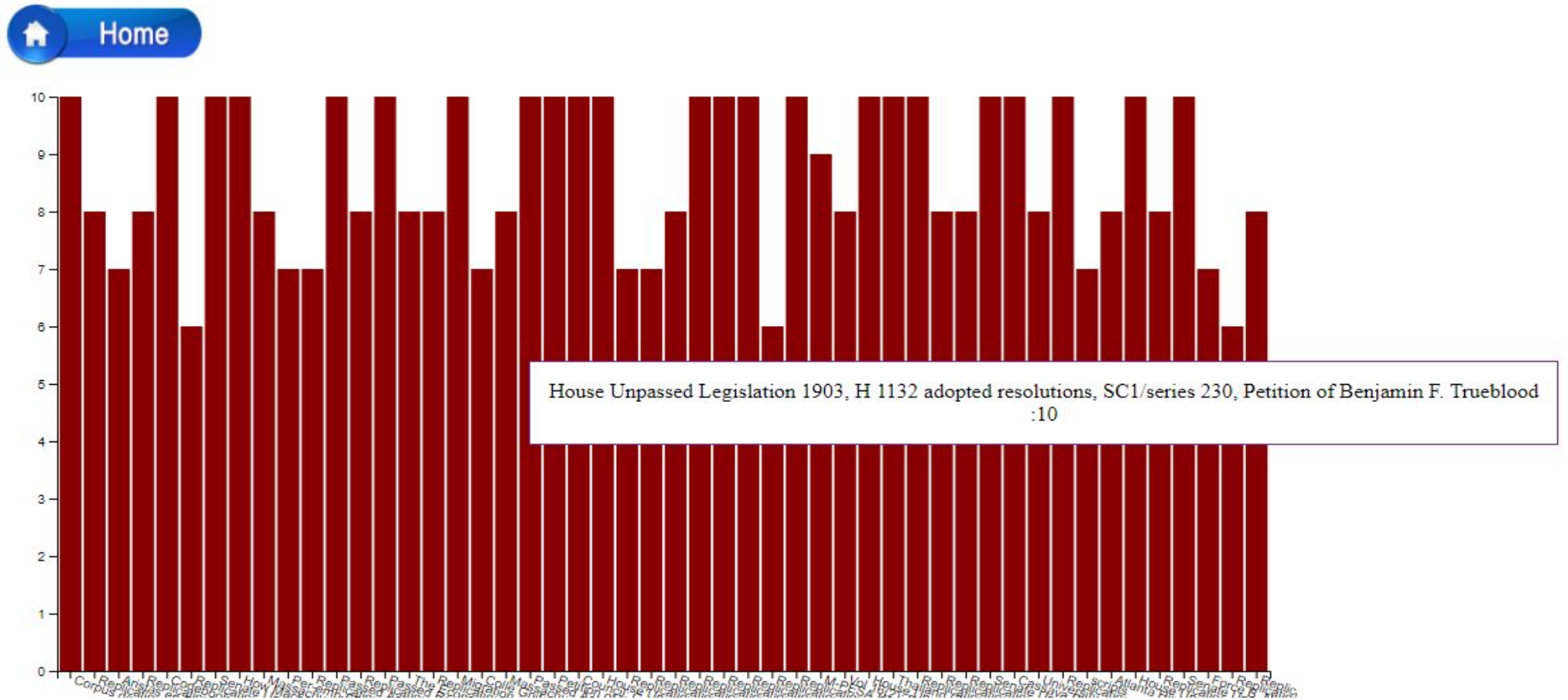
Total maximum score for each dataset is 10.

Dataset Count Based on Data Integrity Score



Sided Bar Charts

Individual Dataset Integrity Score



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 !