Democratizing Data Search and Discovery Platform User Guide*

January 18, 2023

*The Democratizing Data Project has been supported by Schmidt Futures, the Patrick J. McGovern Foundation, the Alfred P. Sloan Foundation, the National Science Foundation (National Center for Science and Engineering Statistics), the Department of Education (National Center for Education Statistics), the US Department of Agriculture (Economic Research Service and National Agricultural Statistics Service) in partnership with New York University, Elsevier, Johns Hopkins University, the University of Texas – Austin, the University of Texas -San Antonio and the University of Maryland.

Contents

1 INTRODUCTION TO THE USER GUIDE

1.1 Overview

The Search and Discovery Platform has been developed to describe how datasets identified by a set of federal agencies have been used. This user guide provides information about the Search and Discovery Platform workflow. The Guide is designed to provide information to agency staff and to researchers who are using the platform to understand how datasets are used and want to know more information about how the reported results were generated. The Guide is also designed to encourage agency and researcher communities to contribute to the platform, and thus increase the value of data for both for themselves and the community at large. Understanding data use and value is a complex endeavor, and the platform will need the contribution of many experts to be fully successful.

With these two goals in mind each chapter unpacks a piece of the workflow: a brief summary is followed by a non-technical description with links to more details for those who are interested. More technical information is provided in the appendices.

The structure of the user guide is as follows. It begins with the vision, goals and context (Chapter 2). Chapter 3 provides a roadmap to each of the subsequent chapters. The **process** for generating the underlying information (the metadata) is described in Chapters 4-6. It begins by describing the source corpus (Chapter 4), the Machine Learning models that are used to find how datasets are used in publications (Chapter 5) and how the output is validated (Chapter 6). The following sections (Chapters 7-8) describe how users can **access** the data through SciServer (Chapter 7) and through the API (Chapter 8). The current **use** of the datasets is through the researcher dashboard (Chapter 9) and the network visualization tools (Chapter 10), although new uses can always be developed by the community through SciServer. The concluding section (Chapter 11) describes the ***potential user community*** and identifies ways in which the agencies can engage with the community. It describes how government agencies and researchers are beginning to use the tools, and provides information about other ways in which stakeholders can become involved - including participating in upcoming workshops, developing better models, contributing new usage measures or providing links to missing documents or data providers. The appendices provide details about the data models and dictionaries.

The project team members would be thrilled to receive suggestions about how to improve the guide, the platform, or any of its components. Please send any suggestions to: info@democratizingdata.ai.

2 BACKGROUND AND CONTEXT

2.1 Vision and Goals

The vision of the Democratizing Data project is to improve the practice of both government policy and research, by providing evidence about how datasets are used and building an ecosystem of agencies, researchers, and other stakeholders that are committed to revealing the value of data and evidence.

The Search and Discovery Platform advances that vision by providing a platform with multiple access modalities - a user dashboard, Jupyter notebooks, and an API -whereby agencies can get a better understanding of how their data are used.

2.2 Agency Questions

Agencies have identified a set of questions: each modality is structured to provide agency staff and researchers a different lens into the questions. It is likely that agencies will develop other questions as the program develops more understanding about how datasets are used, and their impact on evidence and policy, so this list should be seen as a starting not an ending point.

1. **Basic Usage Information**

   How much are agency datasets used in research and how has that usage changed over time? How often is each one of an agency's identified dataset used in research and how has that usage changed over time?

2. **Details about the agency's portfolio**

   What topics are an agency's datasets being used to study and what publications are associated with each topic?

   What topics are each one of the agency's identified datasets used to study in research and what publications are associated with each topic? What other datasets are being used to study each topic?

3. **Drilling into the details for each dataset**

   Who are the main authors using each agency's datasets? Who are the main authors using each specific dataset? What are the publications associated with each author? What institutions are the centers of use for each agency dataset and in what geographic locations are the institutions located?

2.3 Response to Legislative Mandate and Committee Recommendations

The platform directly supports Section 202(c) of the Evidence Act Title 2 (OPEN Government Data Act), namely to

- Facilitate collaboration with non-Government entities (including businesses), researchers, and the public for the purpose of understanding how data users value and use government data
- Engage the public in using public data assets of the agency and encourage collaboration by publishing on the website of the agency, on a regular basis (not less than annually), information on the usage of such assets by non-Government users, and
- Assist the public in expanding the use of public data assets

It also directly supports several of the recommendations to OMB of the Advisory Committee on Data for Evidence Building (see also Figure 1)

Measure and report data value. The production of value (or "utility") is inherent to the core responsibilities of statistical agencies and, as such, is critical for the NSDS. There are several dimensions of value—broadly, adherence to democratic and equitable values and providing value to the public and, more specifically, value of the data assets, value of NSDS capabilities, and value of the data service itself. The NSDS should model an approach to measure and report on the value of each of these aspects, including the following actions:

- ***Produce an NSDS data inventory with usage statistics.*** The NSDS should develop and maintain a publicly available inventory of NSDS data assets in keeping with Evidence Act requirements for agency data inventories. While not a full measure of value, as a baseline, this inventory should include usage statistics. To support a more seamless experience for users, the NSDS data inventory should model the format and content, including detailed metadata, that could be used to harmonize other data inventories and catalogs.
- ***Develop concrete measures of value.*** The NSDS should develop and publish concrete measures of value, including exploring ways to measure the impact and the value of evidence for different stakeholders.

Figure 1: Recommendation of ACDEB

- Measure the utility of data (Recommendation 1.6)
- Optimize user experience for the Standard Application Process (Recommendation 1.10)
- Promote Accessibility and Auditability through data search and discovery Recommendations (5.1 – 5.4)

The platform can also support the Standard Application Process by providing information to researchers about the use of federal data assets in the Federal Statistical Research Data Center system.

2.4 History

The project began in 2016 when NYU was asked by the US Census Bureau to build a secure environment to host confidential microdata to inform the decision making of the Commission on Evidence-Based Policymaking. It quickly became apparent that simply hosting data was insufficient; a search and discovery functionality was needed to find out

(i) **how** data are used, so that government agencies can better understand the utilization of their data investment portfolio and scientists can find other ways in which data have been used

(ii) by **whom**, so that government agencies and their stakeholders as well as scientists (particularly junior faculty and graduate students) can find experts,

and

(iii) to study what **topics** so that government agencies can identify how well their data investments are supporting their mission and scientists can find other work that is complementary to their own.

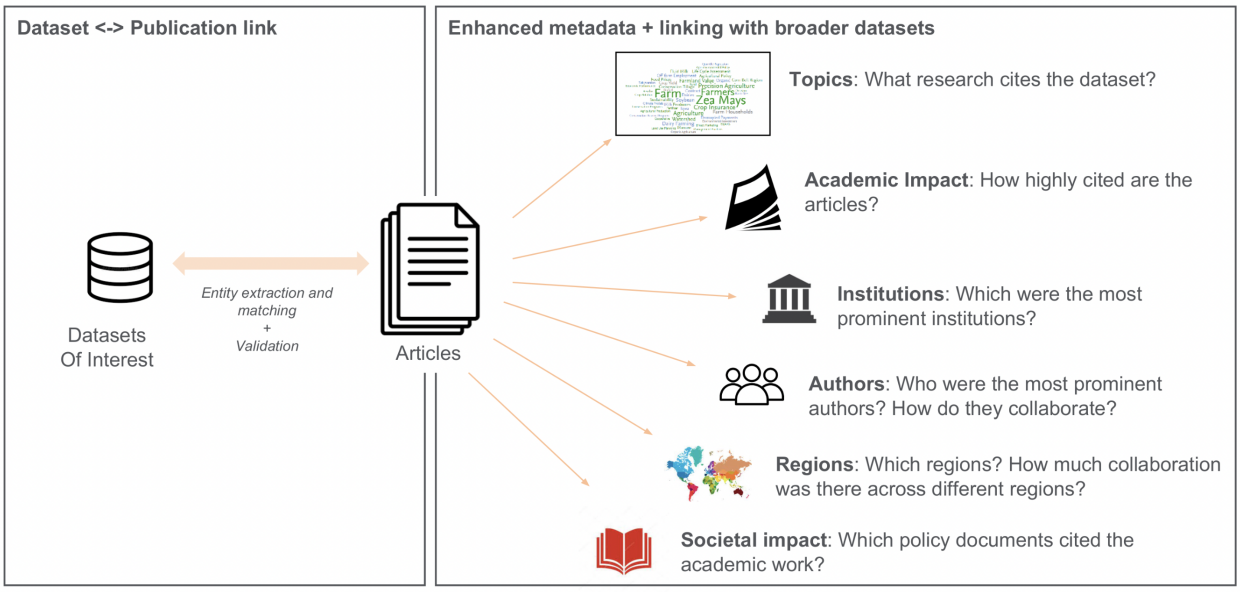


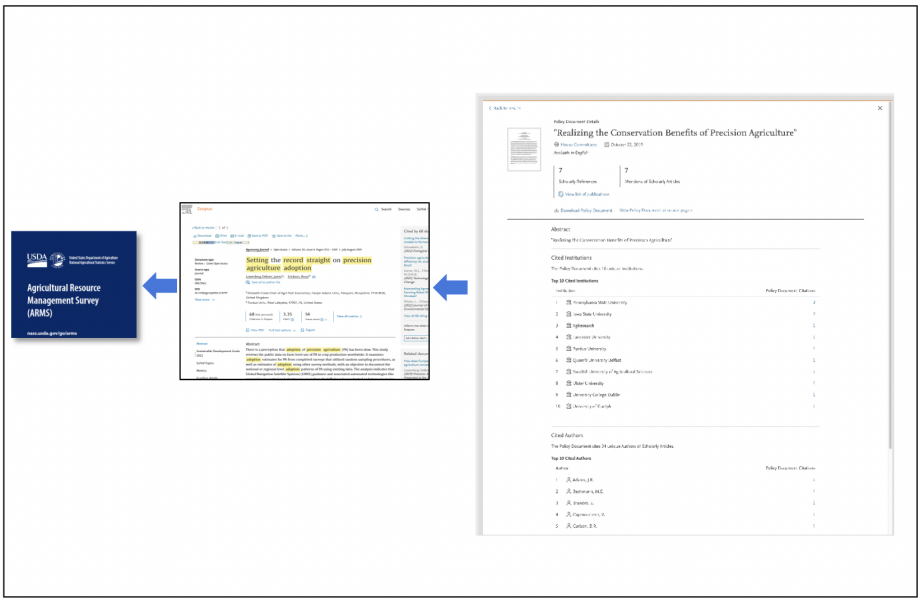
Figure 2: The links between agency datasets and publications

The challenge was that the information about data use is not readily found - A joint project between the developers of Jupyter - Brian Granger and Fernando Perez - and Julia Lane of NYU was funded by the Alfred P. Sloan Foundation and Schmidt Futures to see if it was possible to use Machine Learning and Natural Language Processing tools to automate the discovery of dataset mentions in scientific documents. The team developed and hosted the first rich context competition in 2018; the results were reported in a Sage book published in 2019.[1]

A national conference was subsequently hosted at the National Press Club in November 2019 – also funded by Schmidt Futures and the Alfred P. Sloan Foundation - which was designed to produce a roadmap to identify the opportunities, gaps, and necessary investments, develop an interdisciplinary community of computer scientists, life scientists, and social scientists who can work together to address the problems and engage key stakeholders, notably funding agencies, and government agencies.

The output from the workshop fed into continuing work and during 2021, the effort resulted in a Kaggle competition, known as Show US the Data to develop open algorithms that would improve on the previous efforts. Over 1600 data science teams worldwide competed. The winning algorithms were unveiled at a conference in October 2021 hosted by the Coleridge Initiative and the open access consortium CHORUS and they are the ones currently used in the platform.

A pilot was then initiated with one of the CHORUS board members, Elsevier, to test the possibility of connecting agency datasets with a fully curated corpus of publications, Scopus (Figure 2), and policy documents (Figure 3).



ELSEVIER

Organisation	# documents
Government Publishing Office	17
House Committees	14
Nicholas Institute	7
IFPRI	6
Union of Concerned Scientists	5
World Resources Institute	5
NBER	4
State of Hawaii	3
United States Census Bureau	3
RAND Corporation	3
State of Minnesota	2
Congressional Research Service	2
The American Enterprise Institute	2
Belfer Center	2
Commonwealth of Massachusetts	2
Environmental and Energy Study Institute	2
Third Way	1
National Renewable Energy Laboratory	1
Lincoln Institute	1
CEEPR	1
Information Technology & Innovation Foundation	1
Center for Freedom and Prosperity	1
U.S. Government Accountability Office	1
Senate Committees	1
Wilson Center	1
Brookings Institution	1
RTI International	1
Federal Register	1
Resources for the Future	1

Figure 3: The potential next links to policy impact

2.5 Academic and Research Partners

The project is the result of a collaboration among six academic partners. The role of each is described below.

NYU: New York University (NYU) is responsible for overall coordination and management of the effort to develop and implement the platform. The NYU team is the primary agency contact.

Elsevier: Elsevier provides the information infrastructure - the curated corpus of documents such as publications and patents - that provides the basis on which to run the algorithm, optimizes the search space, applies the existing algorithms to search and discover specific datasets, and produces metadata to feed into the validation process and the API, dashboards, and Jupyter Notebooks.

The Institute for Data Intensive Engineering and Science (IDIES) at Johns Hopkins University (JHU): IDIES ingests and processes the metadata output from the ML algorithm into a database that can be validated in the validation tool. then feeds the validated output to the API. IDIES developed the validation tool so that the agencies or their designated collaborators can validate the output. IDIES also developed SciServer, a collaborative, web-based science platform.

Texas Advanced Computing Center (TACC): Designed, developed, and is implementing the API to disseminate the validated metadata received from JHU. TACC also has enabled the front-end tool implementation by enabling the web connector for Tableau software for visualization in a dashboard. TACC will be developing a browser-based dashboard.

University of Texas at San Antonio (UTSA) is validating the ML output for some agencies. They will also work to support ML model development and the researcher engagement at conferences. Their team will play an important role in analyzing and promoting the progress and success of the

project.

University of Maryland, College Park (UMD) is developing a website that describes the methodology, approach, and relevant materials for the user community. They will jointly host one or more workshops for researchers to react to the findings and measurement.

2.6 Current Agency Partners

National Center for Science and Engineering Statistics (NCSES): NCSES supported the work from the earliest stages. Their goal was to show how NCSES data were being used and tie the information into their revamped website. Later, the development of usage statistics is likely to be important as they establish the National Secure Data Service.

US Department of Agriculture: Economic Research Service (ERS) also supported the work from the earliest stages. It set the initial direction of tool development and implementation. The National Agricultural Statistical Service (NASS) worked with ERS to identify the initial USDA data sets and possible visualization tools, identified validators (either internal or external). They also will conduct research to develop a theory of change and will develop targeted messages to survey respondents to increase overall response rates.

Department of Education: National Center for Education Statics (NCES) has a number of goals:

Relevancy Determine who is using NCES data, including networks of dominant users;

Priority Understand which data sets are being used the most and which the least;

Equity Gain an equity perspective on use of NCES data;

Community Build partnerships with researchers who are conducting research in areas of interest to NCES.

Department of Commerce: The National Oceanic and Atmospheric Administration (NOAA)'s primary objective was to develop a reusable data discovery interface. The overall goal for NOAA is to use machine learning to find citations to NOAA datasets in journal articles and other scientific literature. This knowledge will help highlight the value of NOAA's open data, track the provenance of data used in scientific research, and help new users find trusted data that is relevant to their research topic.

National Institutes of Health: The primary objective is to provide a rich and innovative set of tools for addressing data sharing problems by enabling search and discovery for the subsequent use of NIH funded digital data assets. The NIH project is sponsored by the Office of Data Science Strategy. It will identify a subset of HIV datasets found in Mendeley Data* and find the publications that cite those datasets. It will expand information the use of data in the Mendeley data repository, such as number of views and downloads to include other aggregate measures generated from the Democratizing Data platform such as diversity of use, use within specific disease ecosystems and ecosystem services, use within particular communities, and use of particular historical significance[2] . The team will also develop dataset use at the researcher level, to identify and encourage

*https://data.mendeley.com

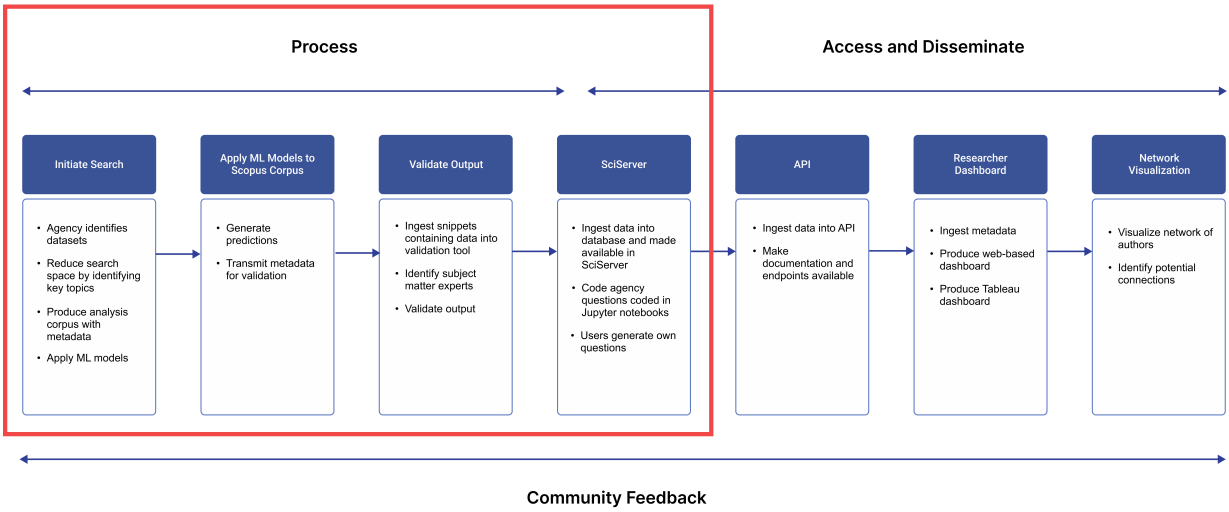


Figure 4: Process

experts who share their expertise, as well as at the institutional level to help inform program investment decisions. The platform will enable NIH/ODSS to extend to and engage with their associated communities, following on from the January 2022 ODSS dataset search workshop.[2]

3 WORKFLOW OVERVIEW

There are three main workflow stages: identifying and finding datasets in publications (process); providing agency and researcher access to information through APIs, Jupyter Notebooks and usage dashboard (access and disseminate), and getting feedback from the user community (feedback).

3.1 Process

Initiate search: An agency identifies a list of datasets of interest (target datasets), ideally including commonly used dataset aliases and, where available, the dataset digital object identifier (DOIs). Elsevier works with NYU to identify a search corpus of full text publications on which the ML models can be run. For more details see Chapter 5.

Apply ML models: Elsevier runs the three ML models to identify candidate datasets mentioned in publications within the search corpus, producing as output:1. the dataset-publication pairs (dyads) 2. snippets of text containing the mention to assist in validation and 3. associated metadata on ML performance and other relevant runtime parameters. Machine Learning algorithms are applied to the licensed full text records in the search corpus to identify those research outputs which contain candidate datasets. The ML algorithms are then further supplemented with matching routines that search for closely related strings (fuzzy matches). The final step in the Elsevier workflow is to generate the metadata for the research outputs identified as containing the datasets. For more details, see Chapter 6.

Validate: The output of those models predicts whether a dataset is mentioned, and Elsevier provides the metadata (journals, authors, institutions, geographic locations and research topics inter alia) associated with those publications to IDIES at JHU for validation.

IDIES retrieves the output information from Elsevier via a predefined secure transport mechanism (in this case delivery of data to an AWS S3 bucket). It then performs a set of automated validations of the format and consistency prior to loading from the file-based format to a relational database. In this process files are cached within the SciServer (described below) environment for further inspection as necessary.

The metadata are ingested into a database (schema is available in Appendix A; the metadata table and data dictionary in Appendix B) and made available to designated agency staff to validate the output in a validation tool. They validate the data by inspecting snippets and validating the ML identification via a web-based tool. This is a first pass in the quality control of the output metadata. There is the potential of adding additional features if resources are available. In particular, researchers could provide additional validation of the results of the ML algorithm by providing feedback on publications that may have been missed or mischaracterized, since the validation process creates rich opportunities for participation and engagement. For more details, see Chapter 7.

Once complete, the full result, including validation is sent to TACC to be ingested in the official API for consumption by end-users.

An administrative dashboard is also being built; that documentation will be made available in Version 2 of this user guide.

The full technical description of this workflow is available in Appendix C.

3.2 Access and Dissemination

There are three access modalities: Jupyter Notebooks, an Application Programming Interface (API), and an interactive usage dashboard. Each is designed to provide different insights into answering the core Agency Questions that are the reason for building the Search and Discovery Platform.

Jupyter Notebooks: The Jupyter Notebooks can be accessed through SciServer `https://www.sciserver.org`, which is a science platform built and supported by the Institute for Data Intensive Engineering and Science (IDIES). Sciserver builds upon and extends the SkyServer system of server-side tools that introduced the astronomical community to SQL (Structured Query Language) and has been providing the Sloan Digital Sky Survey catalog data to the public.

The Jupyter Notebooks are structured to enable researchers to access a database that contains the metadata. Researchers can either program their own queries or access a preprogrammed set of 20 questions about use of the data. SciServer allows researchers to collaborate using hosted data sets in a secure environment, among many other features. For more details, see Chapter 8.

Application Programming Interface (API): The API allows users to download the metadata in order to create their own dashboards or methods of analysis. The endpoints reflect the

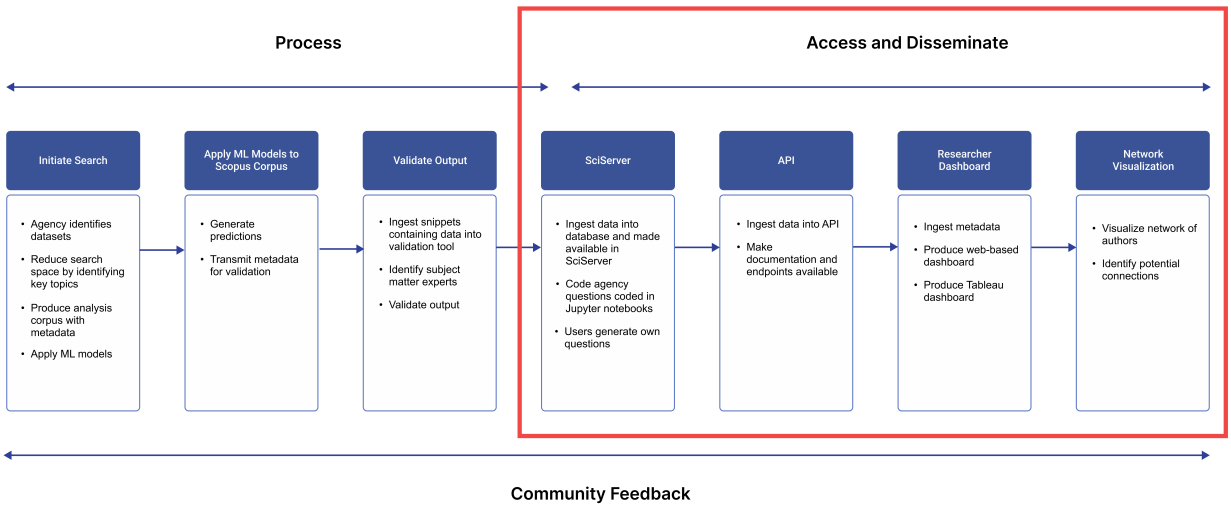


Figure 5: Access and Disseminattion

key information requested by the agencies - publications, citations, authors, institutions, locations, and journals. The API has a web data connector that allows the metadata to be downloaded into Tableau software to create visualizations. For more details, see Chapter 9.

Dashboard: The dashboard and other visualizations illustrate:(1) how data are being used for research; (2) the primary topics of the research; (3) the researchers who have published; (4) how often that research has been cited; and (5) institutions affiliated with the researchers. The dashboards can be structured to inform the agency that is producing the data or also to be a community dashboard that is accessed by the research community and greater public. The dashboard is a type of researcher "leaderboard", identifying the top researchers for different datasets based on the number of their publications and citations. This enables an understanding of who is using the data and their topics of study and helps build collaborative communities. The ancillary benefit to the researchers is a place where their publications can be widely seen and cited in future research, thus encouraging other agencies and collaborators to join in.

In addition, the project team created a sample of five-minute podcasts with top researchers to discover more about their research and how it benefits the public. The researchers were able to provide advice to other researchers using or considering using the data sets, as well as suggestions for how agencies could improve their data and make it more useful. The sample podcasts were published in the Harvard Data Science Review and are available here. As participation increases, these podcasts can be embedded in the API or dashboard, providing practical advice to researchers considering using the identified data sets. For more details, see Chapter 9.

4 CORPUS DEVELOPMENT

4.1 Source Data

The data that Elsevier searches as part of its contribution to Democratizing Data is drawn from Scopus. Scopus is the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings. As at May 2022, Scopus contained 87 million records, from over 7,000 publishers over 100 countries. Around 11,000 new records are indexed every day[3–5] .

Elsevier is itself a significant publisher of research. During 2019, Elsevier accounted for the review, editing and dissemination of 18% of the world's scientific articles. Scopus draws on this rich publishing heritage but also benefits from relationships with other publishers, most of whom have provided licenses that enable Elsevier to search the full text in order to identify key metadata or data elements. Nevertheless, there are some exceptions. For example, for the Democratizing Data project Elsevier is not licensed to undertake full text searchers from Springer Nature so their records must be excluded from the search routines.

4.2 Seed Corpus

The standard corpus creation process used by Elsevier starts with the creation of a seed corpus based upon the target datasets and aliases (or alternative names) for each dataset. In addition, some agencies are able to provide either references to a sample of actual articles in which the datasets have been used or to the names of candidate journals in which articles are likely to be found.

Having high quality information from the relevant agency helps to ensure that the search space is more targeted to the likely publications; this helps improve precision and recall, in particular ensuring that false positives are minimized. The target dataset names and aliases are recorded and form part of the job run metadata.

The seed corpus is created using exact string matching of the target datasets and aliases against Elsevier's Science Direct database. This database contains research outputs published by Elsevier (over 2,500 journal and 40,000 book titles). The Wikipedia entry for Science Direct can be found here.

From the matched publications, Elsevier identifies a set of research Topics. A **Topic is a collection of publications with a common intellectual interest** and can be large or small, new or old, growing or declining. A Topic is defined to be where the direct citation linkages within the Topic are strong and the direct citation linkages outside the Topic are weak. Only the indexed publications are included in Topics. There are 96,000 Topics and these are grouped into 1,500 Topic Clusters. Over time, new Topics will surface, and as Topics are dynamic, they will evolve. For the search corpus creation, the Elsevier team employed the more detailed Topics. A publication can belong to only one Topic and a Topic can belong to one Topic Cluster. More on Topics can be found in the footnote below[†].

[†] https://www.elsevier.com/solutions/scival/features/topic-prominence-in-science
https://www.elsevier.com/__data/assets/pdf_file/0006/548313/Topic-Prominence-Advanced-Webinar.pdf

The Topics, together with any other parameters (e.g., date range) specified by the agency, are then used to identify the corpus of full text articles that form the basis of the Search Corpus. At this point, the publishers of the full text articles are checked and those records are excluded from the corpus where the license agreement does not allow for Elsevier to undertake full text analysis.

4.3 Coverage of Full Text Research Outputs

The following table provides a summary of full text records that are available and searchable based on the calendar year of publication during the period 2017 to 2021. Full text records are, of course, also available for earlier periods although as one goes further back in time, the number of full text records decreases. This table is designed to illustrate the possibilities in the search space for what we have found to be typical periods of interest.

Calendar Year	Scopus Records	Records with Full text	Elsevier's own Full Text Records	Full Text Records Licensed and available for Search	Licensed as % of Full Text Records	Licensed as % of Scopus Records
2021	4,079,818	3,826,784	701,724	3,428,470	89.59%	84.03%
2020	3,856,004	3,583,017	645,188	3,233,503	90.25%	83.86%
2019	3,644,970	3,392,420	597,381	3,091,172	91.12%	84.81%
2018	3,455,000	3,152,839	576,948	2,881,415	91.39%	83.40%
2017	3,270,051	2,843,709	553,854	2,594,161	91.22%	79.33%

Table 1: Full text records available and searchable based on the calendar year of publication during the period 2017 to 2021.

As more fully described in the next chapter, the Machine Learning algorithms are applied to the licensed full text records in the corpus to identify those research outputs which contain candidate datasets. The ML algorithms are then further supplemented with fuzzy matching routines. The final step in the Elsevier workflow is to generate the metadata for the research outputs identified as containing the datasets.

5 THE ML ALGORITHMS

To support the identification of datasets within a set of full text publications, the community was engaged through a Kaggle competition https://www.kaggle.com/competitions in 2021 to develop and identify the best Machine Learning (ML) and Natural Language Processing (NLP) tools. About 1,600 data science teams entered, and seven winners were identified and provided their

data, code and methodology as open-source tools for public use[2,6–8] . Of these seven, Elsevier uses the top three to identify datasets within the full text of the search corpus.

5.1 The Kaggle Models

Model 1 (Deep Learning - Sentence Context)
This model's approach is to use a deep learning-based approach to learn what kind of sentences have references to a dataset. This model takes the longest to run, but also is the most robust to new datasets. It evaluates all of the text within the document.

Model 2 (Deep Learning - Entity names)
This model's approach extracts names of entities from the text and uses a deep learning-based approach to classify an entity as being a dataset or not. This model runs faster than model 1, but is slightly less robust to new datasets.

Model 3 (Pattern Marching)
This model takes a rule-based approach to search for patterns in the document that are similar to a list of existing datasets. This is the fastest model to run, but is the least robust.

During the Kaggle competition, it was determined that the ML models were able to pick up a wider variety of ways in which authors refer to the same datasets than was possible through simple string searches. This was especially true for the winning model "Context Similarity via Deep Metric Learning" which learnt from the context and did not just rely on the way in which the aliases were written in the training set.

In terms of the computer processing, it is simpler and faster to run approaches not relying on Deep Learning methods: Models 2 and 3 performed well if applied to publication domains like the one used on the Kaggle competition. The competition also demonstrated that there was little overlap in the datasets identified by the different models.

Model Name	**Methods Used**
Context Similarity via Deep Metric Learning	Searching candidates in certain format + Filtering out prediction based on keywords + Searching based on the frequency of dataset appearance + Masked language model
Transformer-enhanced Heuristic Search	Searching candidates in certain format + Searching based on the frequency of dataset appearance + Models learning from candidate strings
Simple and Strong Baseline	Searching candidates in certain format + Filtering out prediction based on keywords + Searching based on the frequency of dataset appearance

Table 2: Top ML Models From Kaggle Competition: Models and Methods.

5.2 The Application of the Models

Building on the outcomes and findings of the Kaggle competition, in undertaking the full text search, Elsevier employs all three of the winning Kaggle models. In addition to the identified dataset, each model generates a score that reflects the certainty of the model about the identified mention. The generation of the score is built into each Kaggle algorithm. Elsevier does not apply any thresholds with regard to the Kaggle scores, but rather ingests the full output of datasets generated from the algorithms.

The text identified by the models as being a potential dataset as well as their scores are extracted and stored. In addition, at this point and where licenses allow, a data snippet is generated from the full text showing the text [235 characters] immediately before the candidate dataset text string and the text [235 characters] immediately after the candidate dataset text string. This snippet is used in the validation process i.e. used to identify whether a match is a true match or a false positive.

There is a range of logical possibilities for each full text publication record searched:

- No dataset found in the publication;
- Single dataset found (extracted and single output record produced). This record may or may not be from a target dataset (target dataset in this respect being a dataset provided by the agency or one of its aliases);
- Multiple references to a single dataset found (extracted and multiple output records produced). This record may or may not be from a target dataset(s);
- Single reference to multiple datasets found (extracted and multiple output records produced). This record may or may not be from a target dataset(s);
- Multiple references to multiple datasets found (extracted and multiple records produced). (Again, these may or may not be to target dataset).

5.3 Finding Target Datasets

At this point, the Kaggle algorithms have been applied to the full text and have identified generic datasets. The next step is to identify, from within the Kaggle identified subset of records, the target datasets defined by the client. This is achieved by applying a fuzzy text matching using both the target dataset names and the aliases that have been provided or added.

The fuzzy matching algorithm is an open-source package called FuzzyWuzzy developed in Python. Details about the package can be found here. The fuzziness allows for syntactic differences between the datasets. While running this process it is possible that additional aliases will be found. If that is the case, they are identified and recorded. As with the Kaggle algorithms, a score is generated for each identified pair (i.e., a candidate detection and a target dataset) which is based on the sequences of common characters in both the detection and the target dataset. A threshold is set of the fuzzy scores and only the ones with a score greater than the threshold are kept as a match. The threshold value is determined based on the distribution of the scores across all pairs and the mean character length of the target datasets and aliases. A separate threshold is therefore generated and employed for each of the target datasets in the batch of datasets within the process run.

The fuzzy text matches results in a set of candidate matches or dyads being identified (i.e., where a publication record is linked to a target dataset). The logical possibilities described for the Kaggle algorithms also apply here. Elsevier can generate an output that shows only the target datasets in each publication record or alternatively an output which shows all possible datasets (i.e., including non-target datasets). The core output of this step is thus a results file that for each match found shows: publication ID; target dataset ID; a Kaggle algorithm ID; a Kaggle algorithm score; the data snippet, the candidate dataset text string as found by the fuzzy text matching; and fuzzy text score.

Once the dyads have been identified from the matching process, the required metadata needs to be created. Apart from the data generated through the application of the Kaggle and fuzzy text algorithms, this metadata is generated from the information that is held within Scopus. Scopus includes information that can be used to generate a range of metrics at either the article or journal level, for example the Citescore for a journal, the field weighted citation impact (FWCI) and the number of citations for an article. These metrics are, of course, subject to change over time (e.g., as other research makes reference to an article) and hence the metadata we generate are presented as extant at a specific point in time. Elsevier also provides the relevant research classification information (Research Topics, All Science Journal Classification) for the record. Each metadata field is carefully defined in a data dictionary and in a manner that facilitates subsequent validation checks (e.g., are formats as specified or numbers within the valid range). The metadata information is generated in JSON format to facilitate subsequent automated machine processing including automated checks on the file formats.

5.4 Future Research

The Kaggle competition asked participants to extract dataset mentions from a document. At a high level, the competition asked participants to define a function that did the following:

f(document) = “dataset 1| ... | dataset n”

Where “document” is a JSON-formatted version of the text of the original document. Each of the top three submissions took a unique approach to the competition and offer valuable insight into how to solve this problem. The top two submissions incorporated deep learning-based methods, but the third-placed submission is a rules-based model. The top models all brought their own preprocessing, classification, and post-processing schemes. After the competition, in applying the submissions to new data, some shortcomings of this approach became apparent:

1. Participants didn’t have to offer a confidence value for the detected values. Instead, each model heuristically removes what they considered to be poor submissions.
2. Participants didn’t have to submit where in the document they detected the dataset
3. The constraints on model speed were not tight enough

Elsevier has made a secure environment available to estimate the models in the Elsevier International Centre for the Study of Research (ICSR) Laboratory. The environment is being accessed by team members, and, in future, other researchers to reestimate the ML models. In the short-term, the optimization and improvement of the methods developed via the Kaggle competition is being explored. In particular, the team is reconsidering the problem in the following way. Rather than asking for functions that satisfy the relationship:

f(document) = “dataset 1| ... | dataset n”

we instead ask for functions that satisfy the following relationship:

f(snippet, document) = Pr(snippet contains dataset), dataset token classifications

Where “snippet” is a snippet from the text of the publication which has been identified as possibly containing a dataset reference, and “document” is the entire document. In contrast to the current approach, which asks for a single string as output for an entire document, the proposed approach asks models to produce two outputs for a given snippet and a document. The first is a binary classification for the given snippet and the second is a classification for each token within the snippet.

An example of this format might look like the following:

snippet:

```
For our purposes, the goal is to address causal questions in the context of large-scale
educational assessments (LSAs) . Examples of such LSAs include national surveys such as
the Early Childhood Longitudinal Study (ECLS-K) and the National Assessment of
Educational Progress (NAEP) in the United States, but also cross-national surveys such
as the Organization for Economic Cooperation and Development (OECD)'s Program for
International Student Assessment (PISA) and the International Association for the
Evaluation of Educational Achievement (IEA)'s Program on International Reading Literacy
Study (PIRLS).
```

This snippet would also be paired with the entirety of the text it was taken from. The output of the function given the masked snippet and the document could be the following:

Pr(snippet contains dataset) = 1

for clarity, the tokens that are highlighted below correspond to a 1 and otherwise would be a 0

```
For our purposes, the goal is to address causal questions in the context of large-scale
educational assessments (LSAs) . Examples of such LSAs include national surveys such as
the Early Childhood Longitudinal Study (ECLS-K) and the National Assessment of
Educational Progress (NAEP) in the United States, but also cross-national surveys such
as the Organization for Economic Cooperation and Development (OECD)'s Program for
International Student Assessment (PISA) and the International Association for the
Evaluation of Educational Achievement (IEA)'s Program on International Reading Literacy
Study (PIRLS).
```

This formulation of the problem has a few benefits that address the shortcomings of the current approach:

1. We explicitly ask for binary classification of the entire snippet. By doing this we can leverage the metrics and statistics associated with binary classifiers
2. We ask for the classification of tokens, which helps with the location of valid datasets

3. By including the entire document with the snippet, models that might leverage the entirety of the text (i.e., simple search methods) can be used, but in some way, the model needs to offer a binary classification for both the sentence and the tokens. This presents a fair way to rigorously compare models as the comparison of binary classifiers is well-studied
4. By presenting snippets we can offer the opportunity to generate a balanced dataset rather than what might otherwise be an unbalanced dataset with significantly fewer non-dataset words/tokens than dataset words/tokens
5. This mirrors what we ask the validators to do

The potential benefit is that the revised model would permit agencies to get much better results in two ways. First, predictions about dataset use would have a higher likelihood of being correct (increased precision), reducing the cost of validations. Second, agencies would be less likely to incorrectly reject mentions (increased recall), increasing the quality of information about dataset use.

6 VALIDATION TOOL

Elsevier produces a JSONL-format of the result of the dyad searches and places it in an AWS S3 bucket to which both IDIES and Elsevier have access, as described by the transfer protocol in Figure 4. On a periodic basis, a process downloads data from this bucket, mirrors it to a private location on SciServer, performs a basic set of validations (not to be confused with the human validation that occurs later in the process) to verify that the expected data has been delivered in full and is of the expected quality enabling the continuation of the workflow process - specifically the validations. The data are then loaded into an SQL database of the schema described in Appendix A. Both the database and the original files can be accessed via SciServer (given the appropriate permissions) for further analysis or debugging.

The database (and file mirror) represent another potential access mode in addition to the TACC API, specifically for those interested in working within the SciServer environment.

6.1 Validation Tool Specifics

Each agency has made use of the validation tool to ensure that the ML output is correct. The results:

1. provide insights into how well the models are identifying datasets,
2. find other datasets that are used in conjunctino with the agency datasets to study similar topics, and
3. identify how researchers make use of each agency's data.

The choice of who will validate depends on each agency. Many agencies have asked their own subject matter experts to act as validators. One agency brought in a staff person and made validation one of their tasks. Yet another agency provided funding for a graduate student team from a Hispanic Serving Institution (UT San Antonio) to serve as validators. Those students are

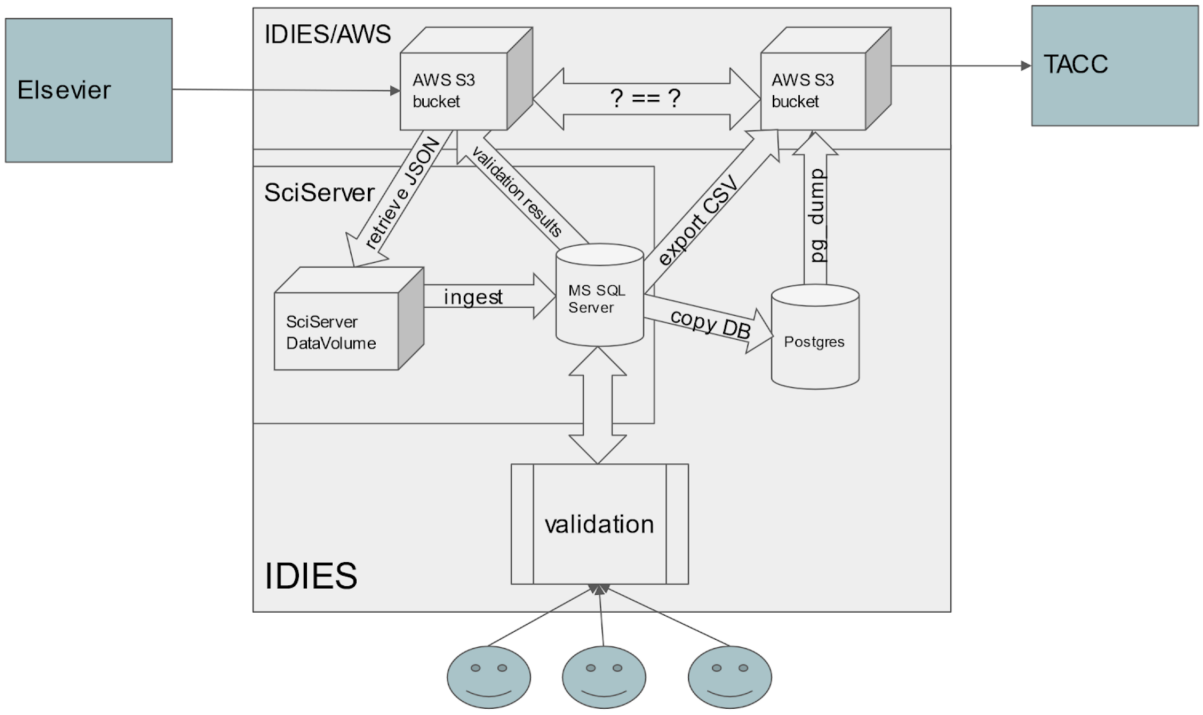


Figure 6: Data Transfer Protocol

writing a report on their lessons learned that they expect to publish in a scholarly journal and will be active in engaging a diverse community in workshops.

The validation tool is very straightforward, and validators have not hitherto needed training in its use. The tool provides reviewers with snippets from actual publications on which the model has been run and which contain references to the dataset being searched for. The snippets contain a candidate phrase identified by the model, and the goal is for the validators to determine if these snippets are referring to the correct dataset.

Registration. Validators can't register themselves in the validation tool. An account is created for them by project personnel, and they will receive their credentials by email. Democratizing Data personnel manage the text snippets that users can review and assign new batches of snippets upon request.

Sign-in. The validation tool landing page has a login form for users to sign in.

Review. When the users log in, they see a bar with the number of snippets assigned and the number of snippets already reviewed. Project personnel initially assign each reviewer a set of snippets to review and can assign additional ones upon request. The reviewers are asked two questions for each snippet:

- **Does the highlighted text in the snippet refer to a dataset?** Confirm whether the machine learning algorithm was correct in considering the phrase as a reference to a dataset in general.

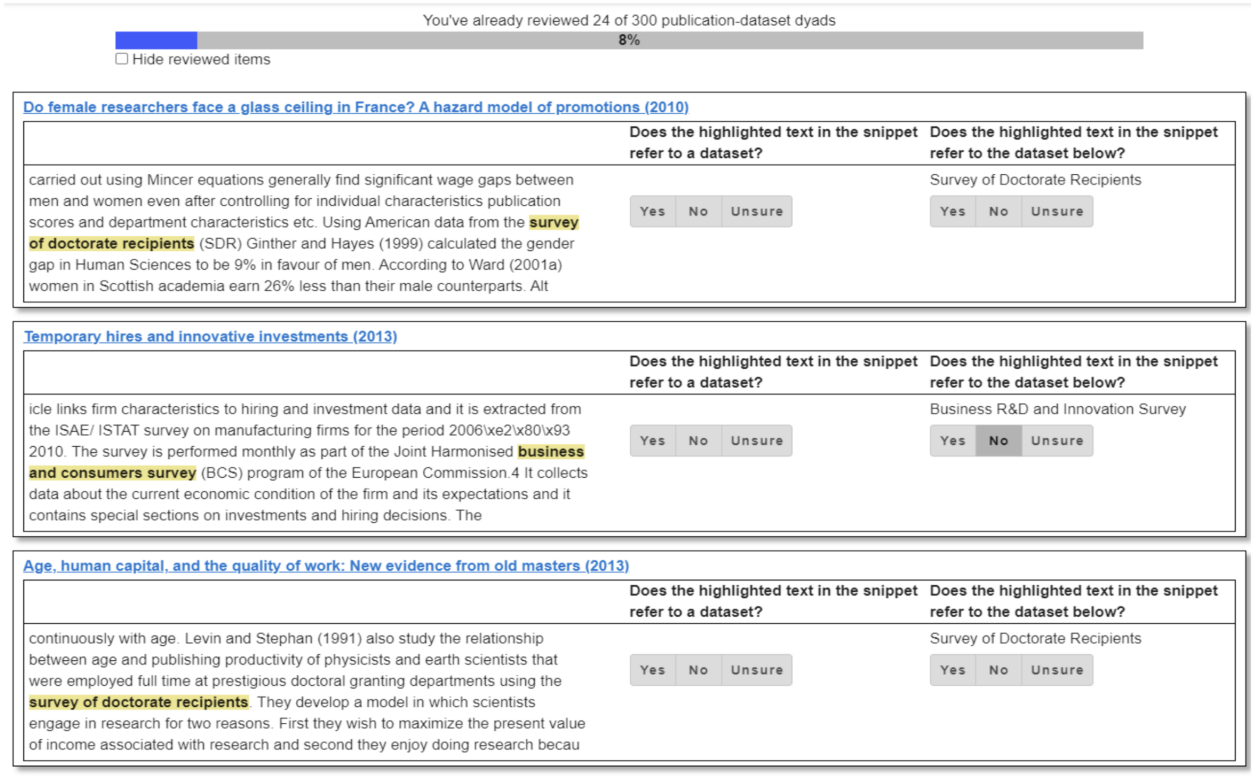


Figure 7: Screenshot of Validation Tool

- **Does the highlighted text in the snippet refer to the dataset below?** Confirm whether the model correctly matched the dataset candidate to one of those dataset names provided by the agencies.

For each question, the reviewer clicks on one box, choosing either yes, no, or unsure. As soon as the users answer the two questions, the text snippet grays out, and they can click the "Modify Review" button to review the text snippet again, changing the answers. Reviewers can also hide/unhide snippets that have already been reviewed by checking/unchecking the checkbox at the top of the page.

At the very bottom of the page, the users can select the number of snippets per page to display. Only ten snippets per page are displayed by default, and up to 50 snippets per page can be displayed.

Users will see a message thanking them when they review all the snippets on the list. Rich context personnel will assign more snippets to review upon request, and users will receive an email notifying them about the new assignment.

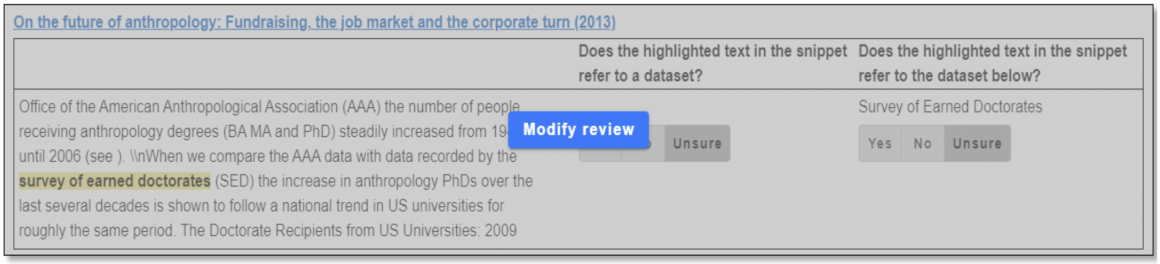


Figure 8: Screenshot of Reviewed Snippet

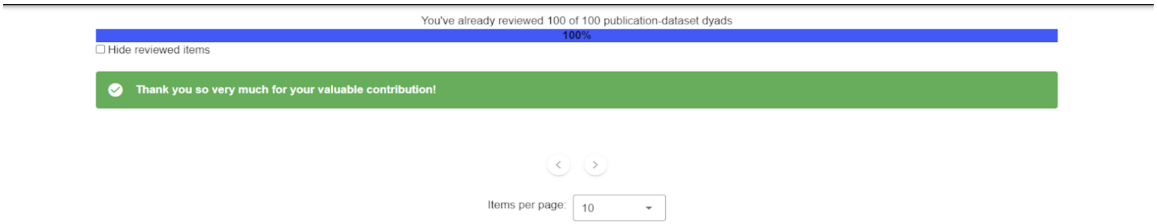


Figure 9: Screenshot of Items per page

7 JUPYTER NOTEBOOKS AND SCISERVER

The Jupyter Notebooks access the database of metadata using an existing successful platform called SciServer. SciServer is built and supported by Johns Hopkins' Institute for Data Intensive Engineering and Science (IDIES) that builds upon and extends the SkyServer system of server-side tools that introduced the astronomical community to SQL (Structured Query Language) and has been providing the Sloan Digital Sky Survey catalog data to the public[‡]. It is particularly appealing because, although it was originally designed to support astronomy research, it expanded to include several research and education tools that made access to hundreds of Terabytes of astronomical data easy and intuitive for researchers, students, and the public[9,10] . For example, one component of the previous system was Galaxy Zoo, a citizen science project that resulted in reliable classifications of hundreds of thousands of galaxy images – a task that was expected to take multiple staff up to 5 years to complete. In the first Galaxy Zoo, more than 40 million classifications were made in approximately 175 days by more than 100,000 volunteers.[11] The current SciServer system has scaled out these tools for multi-science-domain support, applicable to any form of data, including oceanography, mechanical engineering, social sciences, and finance. In addition, SciServer features a learning environment that is being used in K-12 and university education in a variety of contexts, both formal and informal.

[‡]SciServer uses a Docker/VM based architecture to provide interactive and batch mode server-side analysis with scripting languages like Python and R in various environments including Jupyter (notebooks), RStudio and command-line in addition to traditional SQL-based data analysis. Users have access to private file storage as well as personal SQL database space. A flexible resource access control system allows users to share their resources with collaborators, a feature that has also been very useful in classroom environments. All these services, wrapped in a layer of REST APIs, constitute a scalable collaborative data-driven science platform that is attractive to science disciplines beyond astronomy.

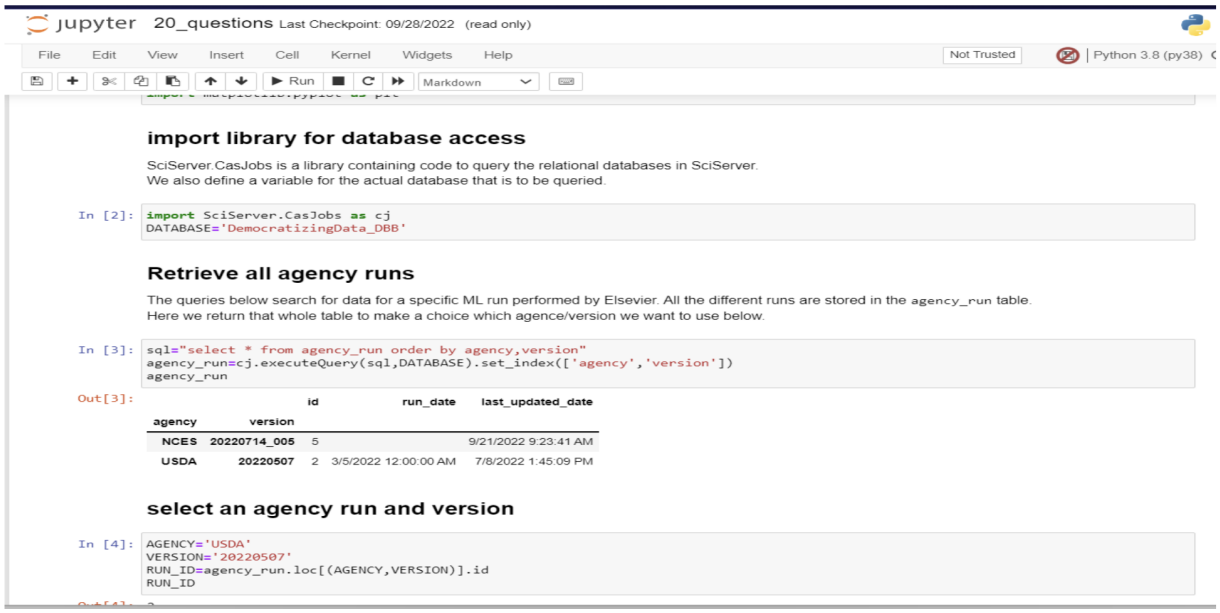


Figure 10: Example of Jupyter Notebook content

7.1 Accessing SciServer

The team has developed a fully-fledged schema and data dictionary and developed Jupyter Notebooks that are accessible through SciServer. Users can first register an account at https://apps.sciserver.org and then contact the team at sciserver-helpdesk@jhu.edu requesting access to Democratizing Data resources, indicating the reasons for the request and their SciServer username.

When access has been granted, example Jupyter notebooks are available on a shared volume alongside the data (see https://apps.sciserver.org/dashboard/files/datavolumes/56 for the base data directory and https://apps.sciserver.org/dashboard/files/datavolumes/56/notebooks for notebooks). The expectation will be that the user is familiar with the use of Jupyter as an interface, has a basic understanding of scripting (typically in Python), and is comfortable with SQL data retrieval.

7.2 Databases on SciServer

With a SciServer account and the appropriate permissions, a user can query Democratizing Data databases via either the CasJobs interface (see https://www.sciserver.org/about/casjobs/) or via the CasJobs Python SDK (see https://www.sciserver.org/docs/sciscript-python/SciServer.html#module-SciServer.CasJobs).

In addition to a master database (ShowUsTheData_v3) which contains records for all the agencies processed, there are databases available for individual agencies (possibly containing multiple so-called "runs") with names of the form "DemocratizingData_{AGENCY_NAME}". These agency-level databases provide data in a manner closer to the API (see Chapter 8), e.g., validated data without licensed snippet information. Individual users may or may not see some or all of these

depending on their access level. For more information on the database schema, please see Appendix A and Appendix B.

8 APPLICATION PROGRAMMING INTERFACE (API)

The Application Programming Interface (API) lets agencies and interested stakeholders be able to programmatically pull the metadata and process/display as they wish. Version 2 of the API is currently available to stakeholders and is updated with agency specific data as it is validated and ready for release.

The API is built using the FastAPI, a Python-based backend which provides routes to database objects with a dynamically generated set of documentation and OpenAPI specifications. The API is containerized and served from Kubernetes and performs calls to functions from the PostgreSQL database. These PostgreSQL function calls optimize query time by making reference to pre-computed database entries in order to decouple query runtime from query complexity. Database updates are propagated to the API via a rollover mechanism that ensures the API will always be performant enough to drive User Interface (UI) applications with minimal downtime. For future capabilities, Fast API can also be integrated with TAPIS (TACC APIs) to provide authentication and fine-grained access controls as needed.

8.1 API Prototype

The prototype Show Us The Data API was a RESTful service that wrapped filter/join functions from the original database schema. The prototype API demonstrated that results from the database could be consumed by external services and clients. The initial prototype was used as an initial springboard for the production API reflected in section 8.2

8.2 API Version 2

Version 2 enables two modes of access to the data set. The first mode supports access for those agency Tableau users who are unable to consume directly from a RESTful API. They can access the data via Web Data Connector and Postgres. The second mode provides access through a RESTful API that supports a generalized Democratizing Data user interface/web client, and provides agencies with the capability of building their own web clients

The Version 2 API includes the following elements:

- TACC-hosted PostgreSQL database that retains production ready, validated data within tables. For complex queries, materialized views are used to maintain query performance;
- Web Data Connect that allows Tableau to retrieve data from the production Postgres Tables and materialized views;
- Production, Pre-production and Development Postgres instances, with access for Democratizing Data to develop and push data directly to the Postgres environment with a rollover-based deployment process for new production data and schema changes;

- (In progress) Automated testing to ensure endpoints and producing proper data when new agency information is pushed to the API;
- (In progress) Landing page for agencies that will answer high level questions and provide further information on how to access additional detailed data.

API endpoints are dependent on what the new data set schema is capable of providing, with requirements for additional endpoints flowing to the Democratizing Data schema team in an iterative development process. The current endpoints are listed in the OpenAPI specification, many of which currently support the "main questions" requirements. These are documented at https://democratizing-data.tacc.utexas.edu/docs as well as here https://prod.democratizing-data.tacc.utexas.edu/redoc and https://prod.democratizing-data.tacc.utexas.edu/docs#

9 USAGE DASHBOARD

Beyond offering an API, the platform currently provides a prototype usage dashboard that enable agencies to visualize answers to agency questions. The initial example for USDA can be viewed here.

As the project evolves, and agencies share resources, it is expected that new dashboards will evolve, as USDA/NASS has shown with their approach here. The proposed production version for a common agency baseline website is in wireframe stage and described in the following sections.

9.1 Basic Usage Information

The initial agency landing page will provide basic usage information, as required by the Foundations of Evidence-based Policymaking Act. Agencies have asked that the landing page provide information about how much agency datasets are used in research and how has that usage changed over time.

9.2 Portfolio Information

Agencies have also asked for reports about their portfolio: i.e., how their data are used. Their questions include: What topics are agency datasets being used to study, and what publications are associated with each topic? What topics are each one of an agency's identified dataset used in research and what publications are associated with each topic? What other datasets are being used to study each topic? Figure 8 shows the answers to those questions.

9.3 Drilling into details

Agencies have also requested information that can also be used by researchers about each dataset in their portfolio, as shown in Figure 9. They have specific questions and the goal of these dashboards is to provide an easy-to-use and clear to understand interface to answer the most pressing questions. Their questions include: who are the main authors using each agency's datasets? Who are the main authors using each dataset, and the associated publications? What

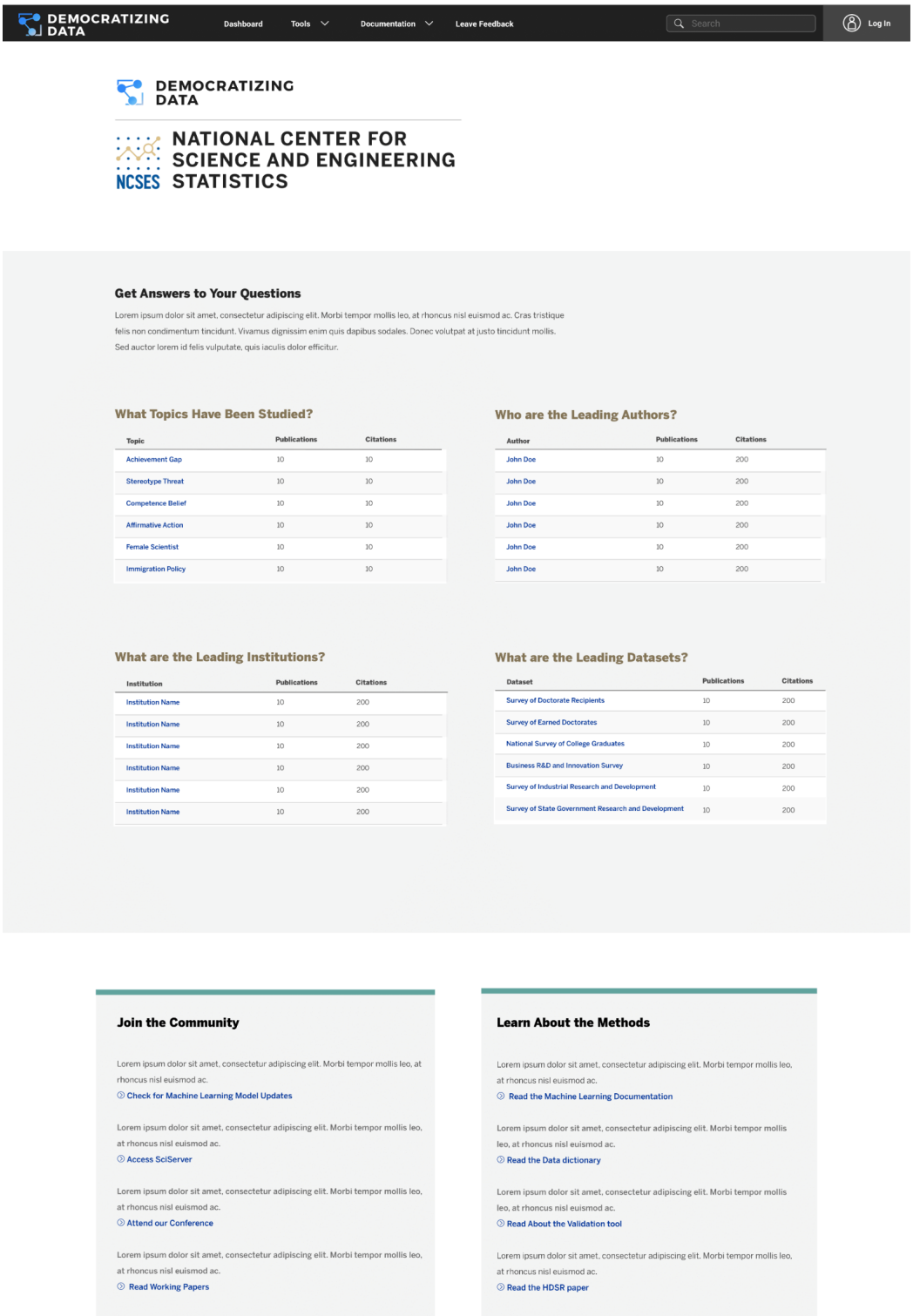


Figure 11: Agency landing page (NCSES example)

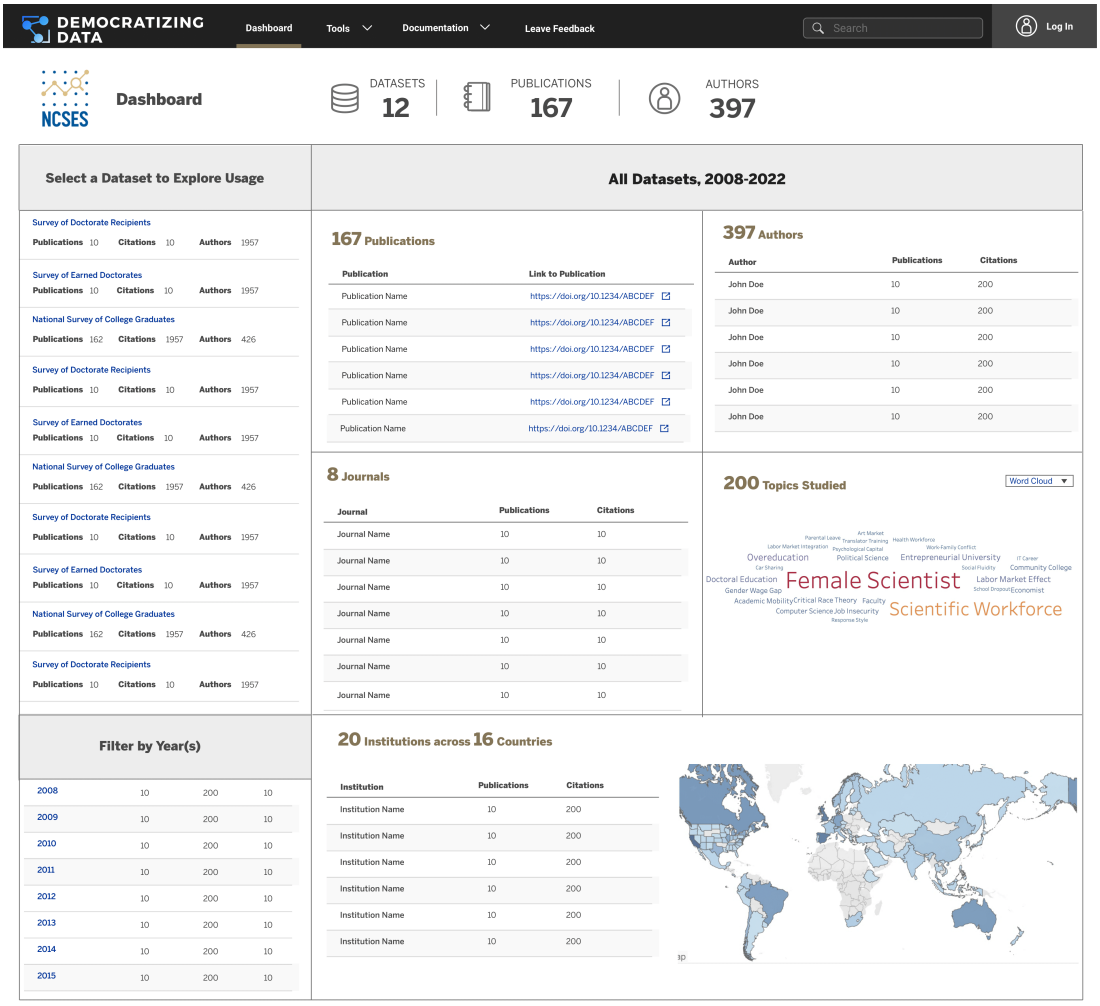


Figure 12: Portfolio information

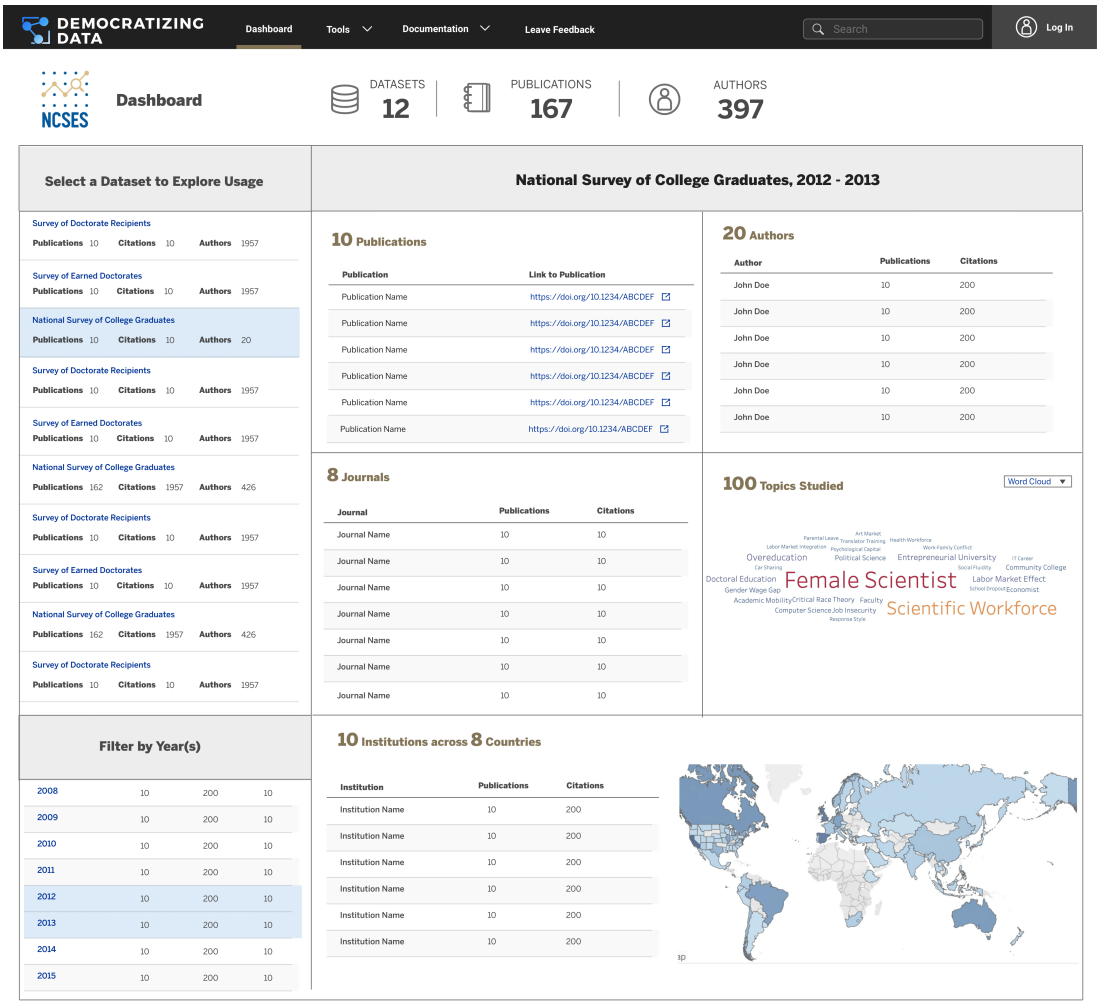


Figure 13: Researcher dashboard with dataset specific information

institutions are the centers of use for each agency dataset and in what geographic locations are the institutions located? What are the journals publishing work using the dataset?

10 DASHBOARD FOR NETWORK EXPLORATION

In addition to the researcher dashboard and the access to the database using the Jupyter Notebooks in SciServer, the team is also developing other user-friendly tools using Jupyter Notebooks. The most developed is the Dashboard for Network Exploration, which has been designed to visualize and explore network graphs derived from the target datasets, and to visualize tables with their associated metadata.

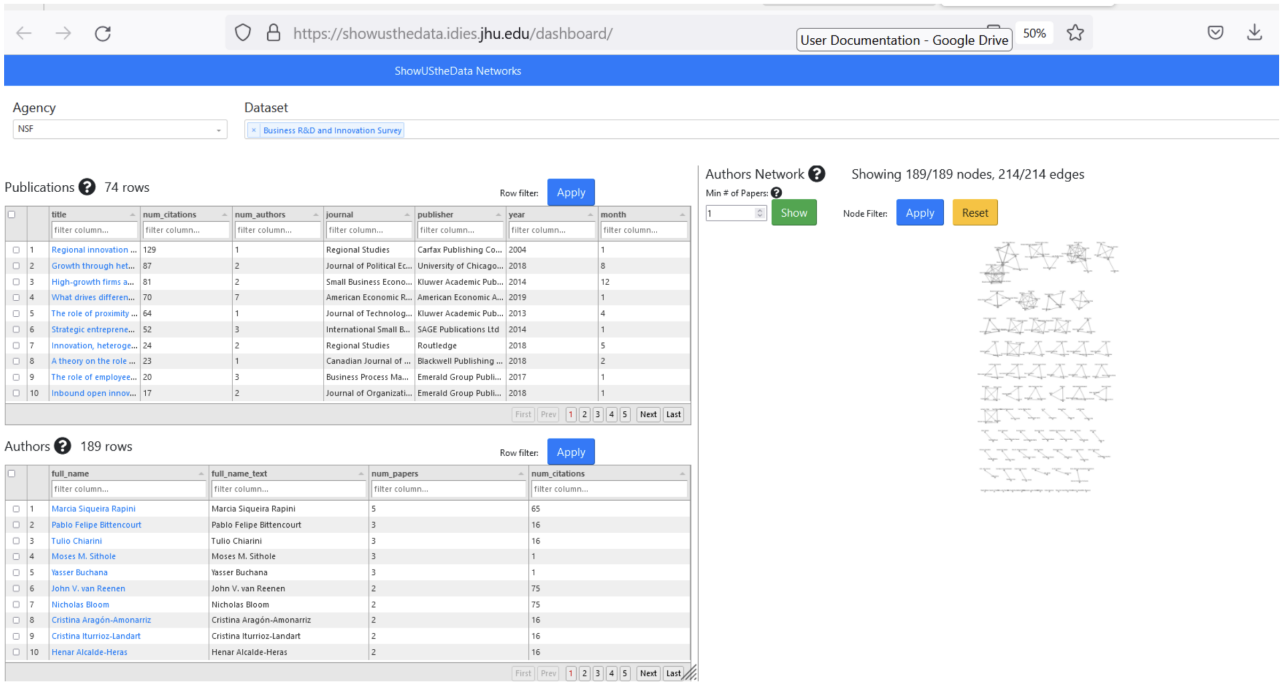


Figure 14: Network Visualization screenshot

10.1 Accessing the Dashboard

The dashboard can be accessed through 2 different interfaces:

- Standalone website: https://showusthedata.idies.jhu.edu/dashboard/ (does not require login).
- Within SciServer in a Jupyter Notebook (requires login).

10.2 Dashboard Filters

The notebooks provide filtering capabilities, where filtering on the network nodes will filter out their associated rows in the metadata tables, and vice-versa.

Author. The nodes of the network are the publication authors.

An edge (link) connecting 2 author nodes is created if those authors appear in the same publication. The size of a network node is proportional to the number of papers published by each author. The width of the edge connecting 2 authors is proportional to the number of papers they have published together.

Institutions. The nodes of the network are the institutions of publication authors. An edge (link) connecting 2 institution nodes is created if an author from one institution appears in the same publication as an author from the other institution. The size of a network node is proportional to the total number of publications associated to the institution. The width of

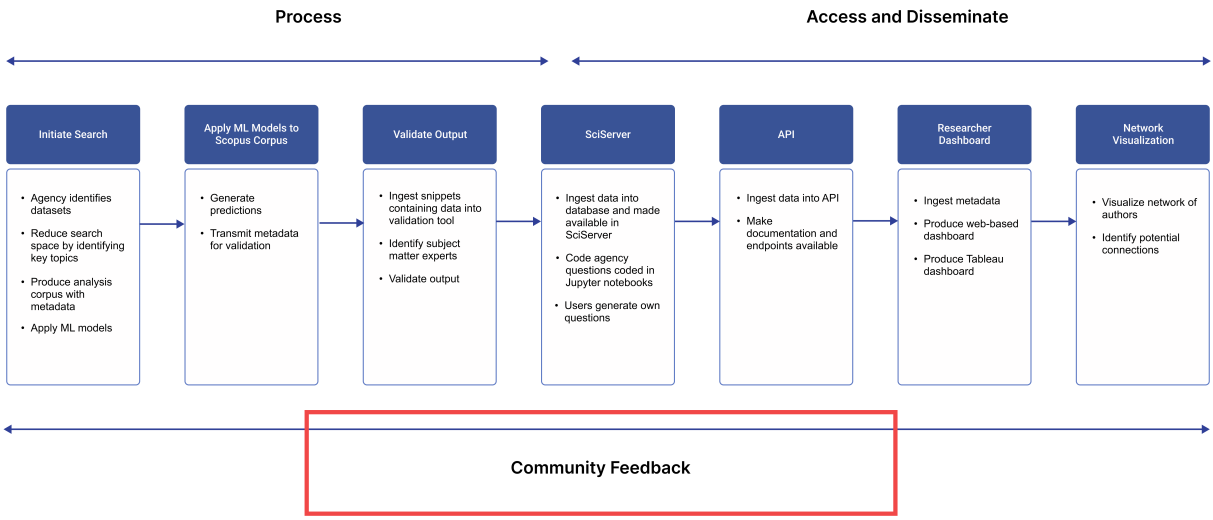


Figure 15: Community Feedback

the edge connecting 2 institutions is proportional to the number of publications their authors have published together.

Topics. The nodes of the network are the topics related to publications. An edge (link) connecting 2 topic nodes is created if the 2 topics appear in the same publication. The size of a topic node is proportional to the total number of times it appears in all publications. The width of the edge connecting 2 topic nodes is proportional to the number of publications the topics have in common.

The following tables and respective columns are shown on the left half of the screen:

- **Publications**: title, journal, publisher, year, month, number of citations and authors. The title column is clickable and opens the journal web page with information about the publication.
- **Authors**: name, number of papers and citations. The name column is clickable and opens the scopus web page with information about the author.
- **Topics**: name, number of papers and citations. Topics are related to the publication, as defined by Elsevier.
- **ASJCs**: name, number of papers and citations. These are ALL Science Journal Classification codes assigned to the publications.
- **Journals**: name, number of papers, and citations.
- **Datasets**: name, number of papers, citations, and authors.
- **Agencies**: name, number of papers, citations, and authors.
- **Institutions**: name, city, country, number of papers, citations, and authors.
- **Countries**: country name, number of papers and citations.
- **Yearly Statistics**: year, number of citations and papers per year.

11 COMMUNITY OUTREACH AND ENGAGEMENT

One of the key goals of the project is to ensure that the user community – both internal to the agency and external – is engaged in substantive ways through workshops, webinars, and a dedicated website that provides opportunity for comment and feedback. Such engagement is not only mandated by the legislation identified in Section 2.3, but also often recommended by National Academies report[§].

While each agency will have its own way of engaging with its internal constituents, there are some common threads that could shape the external engagement with researchers, survey respondents, data users, and specific under-represented groups. However, it is expected that each agency will identify the target user community to participate in the early workshops.

The general format of a user community early workshop would be to explain the platform, provide hands-on experience using Jupyter Notebooks, and to gather feedback on: 1. how to improve the functionality of the platform; 2. usability of the platform; and 3. future possible collaborations between the agency and the user community and within the user community.

Multiple workshops could be structured to serve the different potential constituencies. One might focus on survey respondents, who would react to the usage information in the dashboard. Another might focus on users of the Standard Application Process for the Federal Statistical Research Data Centers. A third might include graduate students, postdocs and other junior scholars who have yet to develop the connections to the empirical knowledge base in a research field.

The workshops will include participation from all the project partners but will primarily be supported by University of Maryland, NYU, and the agencies. The partners are committed to working with the agencies to bring in a diverse and inclusive range of participants, particularly from academic institutions such as Historically Black Colleges and Universities and Hispanic Serving Institutions.

Subsequent workshops could provide input into the theory of change – how investing in data creates value. That theory of change can provide the framework for developing well grounded usage metrics and inform the development of agency questions. As such, a researcher engagement workshop might bring together both active and potential data users, senior and junior researchers interested in the agency mission areas, as well as evaluation experts.

It is also possible that subsequent workshops include the broader federal community. The Evidence Act requires that agencies engage with the user community, and charged three key federal entities with fulfilling that task. These include statistical officials (through the Interagency Committee on Statistical Policy), Chief Data Officers (the Chief Data Officer Council) and Chief Evaluation Officers.

11.1 Other Outreach

Other outreach activities are likely to include presentations at professional conferences of researchers and data users, presentations to federal cross-agency councils, such as the Chief Data Officers Council and the Interagency Council on Statistical Policy, and associations such as the

[§]Such as, for example, the CNSTAT report on "A Vision and Roadmap for Education Statistics"

Figure 16: Error correction

Council of Professional Associations on Federal Statistics. Each agency will be at the center of planning for their outreach activities, with support from the partners.

11.2 Learning from Previous workshops and Outreach

Much previous work can be used in designing the workshops. The Show Us the Data workshop provides a strong basis, since information (reproduced below) was provided by Chief Data Officers, the research community, publishers, and academic institutions.

Chief Data Officer and Evidence Officials

Since the competition focused on uses of data sets in research, the outcomes were most immediately appliable to agencies with scientific mission components. CDOs from agencies for which discovery activities occurred in the competition were invited to review results in one-on-one sessions and then to attend this panel discussion. The Agencies represented in the discussion were Commerce (NOAA), NSF, USDA, Transportation. Given that the breadth of data work in an Agency may cross many mission teams, some agencies had multiple team members participate in the discussion session Their detailed responses are summarized in Appendix D: CDOS.

Researchers

A set of academic and agency researchers were asked a series of structured questions including: (1) how they might use the tools to advance their research; (2) how the tools might advance the work of junior researchers; (3) how the tools might inspire researchers to do their work differently; and (4) how might the researcher community become engaged in this effort? Their detailed responses are summarized in Appendix D: Researchers.

Academic Institutions

Several benefits for researchers at institutions included improved discovery of what data exist and are available, better access to data, and opportunities for collaboration, especially across

disciplines. More use of the data would also create motivation to improve the metadata, e.g., developing and conforming to metadata and citation standards and making sure data are complete. This would also help improve existing governance structures and help integration across existing infrastructures. Institutions want to understand usage and improve discovery and access from their data repositories. Institutions also use a lot of state and other data, so there could be wider applications beyond federal data. Detailed responses are available in Appendix D: Institutions.

Publishers

The workshop participants were asked structured questions to get feedback on what the publisher stakeholder community thinks about the potential of the Rich Context Content project and its machine learning and natural language processing components. The questions related to: (1) Concerns about the Machine Learning / Natural Language Processing (ML/NLP) approach to capturing data use; (2) Additional functionality that would be useful; (3) the value proposition for publishers to participate; (4) How publishers could participate; and (5) where should the application reside and be managed? Their responses are summarized in Appendix D: Publishers.

11.3 Maturity

As the project matures, it is hoped that the community will provide additional input. Most immediately, the community should provide input into the theory of change and what measures should be used to measure the value of data. In subsequent activities, the community could also support the development of the broader information infrastructure. This would include improving the ML tools themselves, filling gaps in the corpus that the ML models missed, and incentivizing both data users and producers to contribute documentation, code, and analytical uses to the platform.

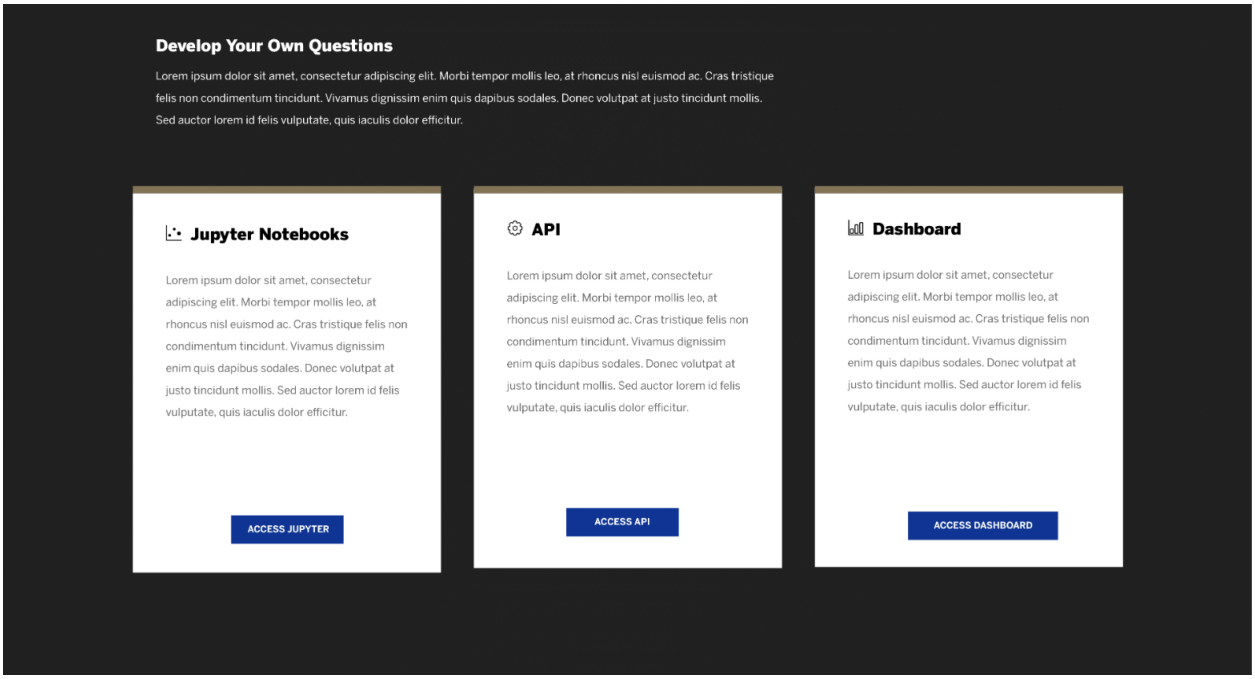


Figure 17: User engagement

Initially the engagement mechanism will be workshops and webinars. However, it is expected that the platform will include substantial modalities that allow for human-computer interaction and error correction .

11.4 Input on User Tools

The initial “ask” will be to get internal staff and researchers to comment on the user tools, their functionality, and the usage measures through staff briefings using the provided tools – the Jupyter Notebooks, the API and the usage/researcher dashboard – as well as develop their own questions.

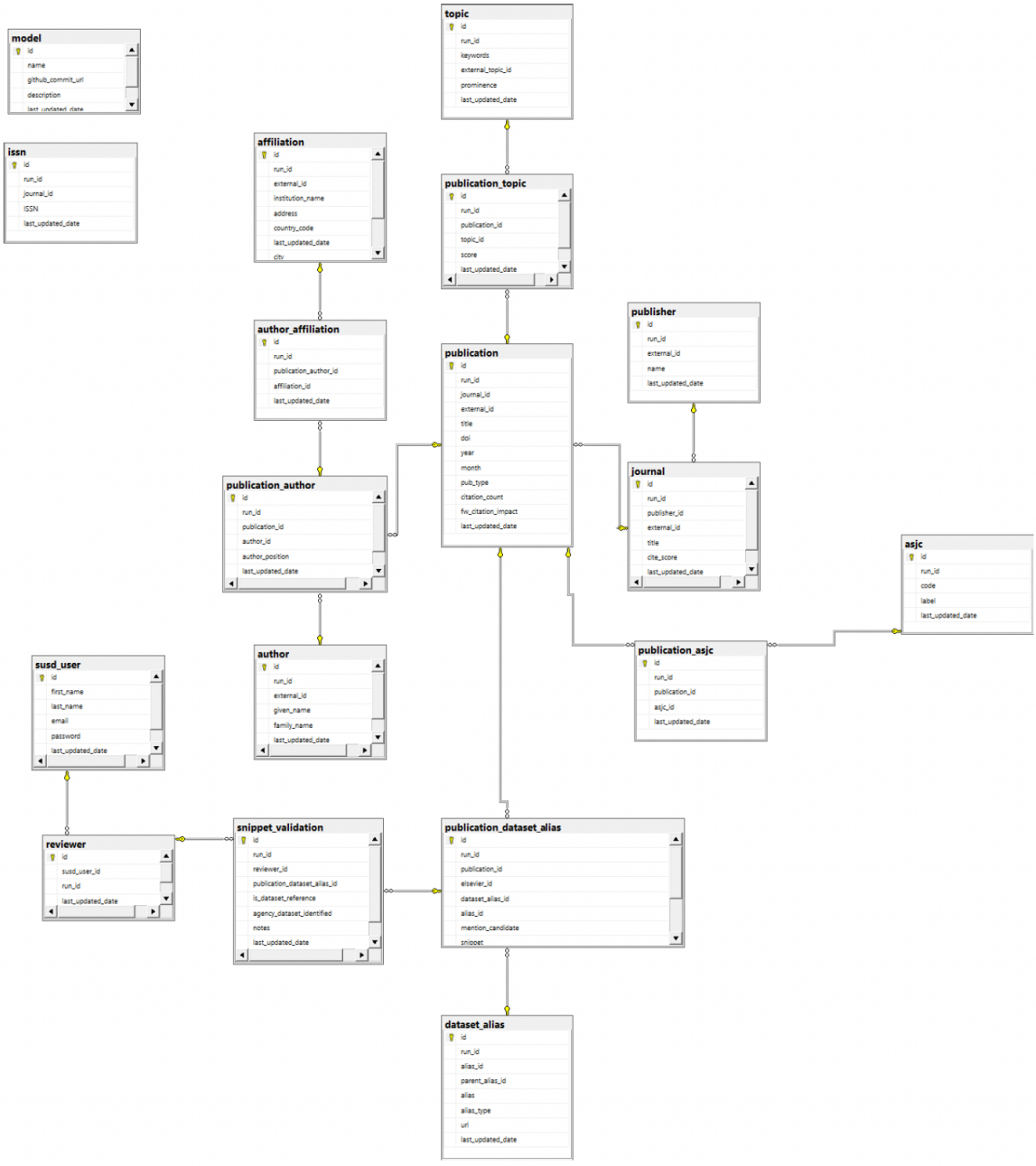
Initially, these will be small workshops focused on training, exploring the platform, and collecting user experience data. During these early workshops, usability testing will also be taking place to inform and improve the tool set.

11.5 Measures Based on a Theory of Change

Agencies have identified a framework that they would like tested with their internal and external communities.

APPENDIX A: METADATA SCHEMA

This describes the metadata schema. The Data Dictionary is provided in Appendix



APPENDIX B: METADATA TABLE AND DATA DICTIONARY

AffiliationGeo : Geo information about affiliations				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
affiliation_id	Nan	bigint	0	YES
q_in	Nan	varchar	-1	YES
q_final	Nan	varchar	-1	YES
nattempt	Nan	smallint	0	YES
boundingbox	Nan	varchar	-1	YES
lat	Nan	float	0	YES
lon	Nan	float	0	YES
display_name	Nan	varchar	-1	YES
importance	Nan	float	0	YES

agency_run: The table with runs for the different agencies				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier to the agency_run table	bigint	0	NO
agency	name of the agency for which the run was performed	varchar	32	NO
version	version of the run for the agency. allows multiple versions on the same data sets, or possibly new runs for the same agency but with different input data sets.	varchar	32	NO
run_date	approximate date the run was performed.	date	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

asjc: All Science Journal Classification code defines the research area of a journal and the articles it contains.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
code	The All Science Journal Classification code which defines the research area of a journal and the articles it contains. There may be more than one ASJC code for each journal / publication. The 334 codes are used here providing a relatively precise definition of research area.	bigint	0	NO
label	TBD	nvarchar	-1	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

author: table with author information				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
external_id	id assigned by Elsevier to the author. allow different authors to be identified across publications, even if they have different names there.	varchar	128	YES
given_name	the unique given name of the author as determined by Elsevier	nvarchar	150	YES
family_name	the unique family name of the author as determined by Elsevier	nvarchar	150	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

author_affiliation: table linking authors to their affiliations in a publication				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
Id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
publication_author_id	identifies the publication_author here linked to publication affiliation	bigint	0	YES
publication_affiliation_id	identifies the publicaiton_affiliation entry here linked to a publication author.	bigint	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	date	0	NO

dataset_alias: Datasets provided by an agency for a particular run and possible aliases				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
alias_id	the alias_id as provided by Elsevier.	bigint	0	NO
parent_alias_id	identifies the parent dataset entry in this table, as identified by the alias_id!	bigint	0	YES
alias	the name of the data set or the alias depending on whether alias_id==parent_alias_id or not.	varchar	160	YES
alias_type	flaf indicating whether the row stores the dataset itself, or an alias.	varchar	50	YES
url	URL to information about the dataset	varchar	2048	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

dyad: The core table with dyads representing dataset references.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
publication_id	foreign key the publication table's id column, identifying the publication within which the dataset reference represented by this dyad was identified.	bigint	0	NO
elsevier_id	REMOVE	int	0	NO
dataset_alias_id	foreign key to the dataset_alias table's id column identifying the match made between this dyad and a dataset alias provided by an agency. if no such match was found this column has a NULL	bigint	0	YES
alias_id	the intrinsic id assigned by Elsevier to the dataset alias, corresponding to the alias_id column in the dataset_alias table.	bigint	0	YES
mention_candidate	the phrase in the publication that was deemed by the algorithm to reflect a reference to a dataset	varchar	1028	NO
snippet	snippet of text surrounding the mention_candidate, meant to provide contextual information to reviewers/validators	varchar	-1	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO
is_fuzzy	column has value 1 if the matching between mention candidate and dataset alias was performed using a fuzuuy algorithm, 0 otherwise.	bit	0	YES
fuzzy_score	in case the matching between mention candidate and dataset alias was perfomed using a fuzzy algorithm, this column stores the score indicating how certain the match was deemed to be.	real	0	YES

dyad_model: model scores for particular entries in the dyad table				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
dyad_id	identifies the dyad	bigint	0	NO
model_id	identifies the model	bigint	0	NO
score	the score of this model for the dyad	real	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

issn: The ISSN / ISBN codes for the journal.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
journal_id	foreign key to the journal table's id column, identifying the journal for this ISSN	bigint	0	YES
ISSN	The ISSN / ISBN codes for the referenced journal/source	varchar	13	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

journal: Journal that a publication in the publication table appeared in.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	Unique identifier and primary key for this table.	bigint	0	NO
run_id	foreign key, identifier of the agency run for which this entry was determined.	bigint	0	NO
publisher_id	foreign key to the publisher table, identifying the publisher for this journal at the time the agency run was executed.	bigint	0	YES
external_id	The Scopus ID for the journal / source	varchar	128	YES
title	The name of the journal / source that the publication was published in.	varchar	1028	NO
cite_score	Citescore is an Elsevier derived metric that measures the relative standing of a journal.	decimal	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

model: The Kaggle models that are run.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
name	the name of the model	varchar	32	NO
github_commit_url	the github url where the commit for this model can be found	varchar	1024	YES
description	description of the model	nvarchar	-1	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

publication: publications discovered in a run				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
journal_id	foreign key to the journal for this publicaiton	bigint	0	YES
external_id	the scopus ID (for Elsevier publications) of this publication.	varchar	128	YES
title	title of the publication	varchar	400	YES
doi	DOI of the publication	varchar	80	YES
year	The year that the publication was published as recorded in Scopus	int	0	YES
month	The month of publication. May not be available. This will be an integer value i.e. 1 = January etc	int	0	YES
pub_type	The type of publication. Types includes - Article, review, book, book chapter, letter.	varchar	30	YES
citation_count	The number of times this publication is cited in Scopus	int	0	YES
fw_citation_impact	The Field Weighted Citation Impact (FWCI) for the publication. This is a measure for how impactful or important a publication is as measured through normalised citations. The number of times cited divided by the expected number of citations of articles in the same year, subject and publication type. World average across papers is 1.0 for this metric. This metric also changes over time.	float	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

publication_affiliation: The table with affiliations on a publication.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier of the affiliation table	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
external_id	id assigned by Elsevier to this affiliation	varchar	128	YES
institution_name	the name of the institution to which a author was associated.	nvarchar	750	YES
address	the address of the author, most likely that of their institution	nvarchar	750	YES
country_code	the three letter country code associated to the affiliaiton, most likely of the institution	nvarchar	10	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO
city	the city for the addres or institute in this affiliation	nvarchar	128	YES
state	if appropriate, the state for the addres or institute in this affiliation	nvarchar	128	YES
postal_code	if appropriate, the postal code for this affiliation	nvarchar	64	YES

publication_asjc: Associative table linking a publication to ASJC entries.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
publication_id	foreign key to publication table's id column, identifying the publication in this relation	bigint	0	NO
asjc_id	foreign key toi the ASJC	bigint	0	NO
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

publication_author: Associative table linking publication and author tables.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
publication_id	foreign key to the publication for this author	bigint	0	NO
author_id	foreigj key to the table with scopus author entries	bigint	0	NO
author_position	position of author in the list of authors on the publication	int	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

publication_topic: identifying the topic assigned to a publication				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
publication_id	foreign key to the topic table's id column identifying the publication in this relation between publications and topics.	bigint	0	NO
topic_id	foreign key to the topic table's id column identifying the topic in this relation between publications and topics.	bigint	0	NO
score	TBD	real	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

publisher: Publishers of the journals the publications were published in.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
external_id	external identifier of this publisher in elseviers scopus repository	nvarchar	128	YES
name	name of the publisher	nvarchar	120	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

reviewer: Reviewers are assigned to validate dyads in the publication_dataset_alias table.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
susd_user_id	foreign key to the susd_user table's id columns, identifying the user corresponding to this reviewer	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

snippet_validation: the validation results for dyads provided by reviewers				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
reviewer_id	foreign key to the reviewer table's id column, identifying the reviewer assigned to validate this snippet	bigint	0	NO
dyad_id	foreign key to the dyad table's id column, identifying the dyad that is being validated.	bigint	0	NO
is_dataset_reference	if the value in this column is 1 it indcates the dyad indeed has identified a reference to a datasdet, if 0 it is not a dataset reference, if -1 the reviewer was unsure about it.	smallint	0	YES
agency_dataset_identified	if the value in this column is 1 it indicates the dyad indeed identified the specific dataset provided by the agency, if 0 it is not a reference to that dataset, if -1 the reviewer was unsure about it.	smallint	0	YES
notes	any notes the reviewer attached to the dyad being reviewed	nvarchar	-1	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

susd_user: A user of the validation tool.				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier of this table	bigint	0	NO
first_name	First name of the individual.	varchar	100	YES
last_name	Surname of the individual.	varchar	100	YES
email	email of the user.	varchar	100	YES
password	encrypted password of the user.	varchar	100	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

topic: topics defined by Elsevier and assigned to publications. consist of three concatenated keywords				
Column name	**Description**	**Data type**	**Length**	**Is nullable**
id	unique identifier for this entry	bigint	0	NO
run_id	identifies the agency run for which this entry was determined, foreign key to agency_run.id	bigint	0	NO
keywords	a topic is defined in Elsevier by three keywords. This columns sotres these as a —-separated string.	varchar	1028	YES
external_topic_id	external identifier of this topic provided by Elsevier.	varchar	128	YES
prominence	TBD	real	0	YES
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO
last_updated_date	last time the row was updated. generally the time of creation of the row.	datetime	0	NO

APPENDIX C: TECHNICAL WORKFLOW DESCRIPTION

Partners

EL Elsevier

IDIES Institute for Data Intensive Engineering and Science, Johns Hopkins University

NYU New York University

TACC Texas Advanced Computing Center at the University of Texas at Austin

Roles

SYSADMIN a defined user of the admin dashboard who has been authorized by the Democratizing Data leadership to initialize projects that an agency has requested and to assign users as administrators of those projects

ADMIN a defined user from the agency staff who has been assigned by the agency to manage a project, configure its input parameters, monitor progress, and assign reviewers, etc.

REVIEWER a user assigned by the admin to review/validate the results of the machine learning algorithms.

Glossary/Definitions

Project A dataset search and discovery activity requested and defined by a funding entity (currently federal agency). It is characterized in a statement of work by a set of datasets that are targets of the inquiry, as well as a set of restrictive parameters.

Main Alias The main alias is the dataset name that most commonly describes the dataset and/or the dataset which the agency would wish results to be grouped by.

Alias Type Many Datasets have short form names or alternative names that are used instead of the Main Alias. These are dataset aliases. A specific form of alias is an acronym or abbreviation. Where such aliases exist they can form part of the search routines.

Research Output A document / publication representing the formal results of research. Research outputs include journal articles, conference papers, review papers, book chapters, books etc.

Steps

Indicated by a sequence number and by location

Step 0: Project initialization [IDIES]

A **SYSADMIN** initializes a new project in the admin dashboard.

This requires the following actions:

1. Adding project level metadata such as:

- Formal department name (e.g., USDA or US Department of Agriculture)
- Formal agency name (e.g., NASS or National Agricultural Statistical Service)
- Unique identifier which includes a date stamp and version number (e.g., 2022_12_15_v1)

2. Assigning a "defined user" to be ad ADMIN for this project and communicating with that ADMIN user to explain entry to the system.

Side effects:

- An entry representing the project will be written in the `agency_run` table.
- An entry representing the agency **ADMIN** will be added to the `reviewer` table.
  - if no user exists yet an entry will also be added to the `susd_user` table.
- An entry will be added to the `agency_run_history` table.

Step 1: Project definition [IDIES]

Agency ADMIN uploads target dataset-alias file format: either

csv[¶]:

Each row should correspond to a dataset name/alias
columns:

- `Main_alias_id`: identifier, created by NYU upon receipt and unique in file
- `Main_alias`: the string identifying name of dataset identified by the agency as the formal name of the dataset
- `_alias_name`: identifies and aliases that are commonly associated with `Main_alias`
- `alias_type`: one from [`main_alias`, acronym]
- Dataset DOI: if available

or

JSON:

```
{"dataset": <dsname>,
 "aliases": [{"alias": <alias>,
              "type": <one of ["alias", "acronym"]>}

 ],
 ...
}
```

ADMIN set other search config parameters:

- data range: [start-date, end-date] where data is ISO-8601 compliant (YYYY-MM-DD) and based on calendar year
- US author flag: boolean

[¶]Note that Excel is discouraged because it typically eliminates leading zeroes and can thus negatively affect the functionality of identifiers

ADMIN clicks button to save or submit:

- **save**: result stored in database
  - Data model extension needed.
  - History of consecutive save-s stored.
- **submit**: Elsevier notified
  - TBD how: likely file uploaded to new S3 folder that **Elsevier** is listening to
  - milestone noted in database

Step 2: Identify relevant topics for search corpus creation [EL]

Elsevier uses the job identifier created in step 0 to organize its work and undertakes process as follows:

- Take datasets names and aliases from step 1;
- Exact text marching on Science Direct;
- Identify the Research Topics on the matches generated;
- Aggregate the counts of matched research outputs by Topics;
- Apply filter that excludes those topics with less than a count of 5 research outputs.

Elsevier transmits a JSON file containing resulting list of Research Topics from Science Direct to **IDIES** with the following aggregate metadata:

- unique run identifier;
- Research Topics;
- count of research outputs against the filtered Research Topics.

NYU and IDIES review and select the research topics with EL.

IDIES shows the result in the admin dashboard and stores milstones in database.

ADMIN inspects result (Note that in V2 agencies will be able to provide input).

Step 3: Determination search corpus [Elsevier]

Elsevier determines the input for the ML Algorithms:

- Search Scopus using Research Topics and the search config parameters and identify the research outputs records that are theoretically available for the search corpus.
  - Output 2.1 Number of records theoretically available in Scopus.
- Filter the records to exclude those for which full text does not exist.
  - Output 2.2 Number of records for which full text exists.

- Exclude records for which full text search is not allowed because of licensing agreements.
  - Output 2.3 Number of records for which full text exists and we are licensed that may be searched.
- From this filtered set of research outputs, identify the Research Topics and calculate the number of research outputs that are linked to each topic.
- Filters topics to include only those with more than 5 research outputs per topic.
  - Output 2.4: List of research topics with counts of research outputs greater than 5

Elsevier transfers Outputs 2.1 to 2.4 to **IDIES** in a JSON file.

IDIES displays results on the dashboard and stores milestone in the database.

ADMIN inspects the result (Version 2 will allow more interaction).

Step 4: Running ML algorithms [Elsevier]

Elsevier runs the ML models on the Search Corpus:

- Record the results of the models in a way that enables the different results generated by the different models to be later compared (i.e. which datasets were found by which models).
- Perform fuzzy match using the alias names to determine whether the datasets found by the ML models ae on the agency dataset list and tags with the unique identifier.
- Filter the set of matched records to indicate which are associated with target datasets and those where snippet cannot be generated for licensing reasons

Elsevier provides aggregate metadata about the run to **IDIES**.

For target dataset:

- **Dataset name;**
- **Unique ID for dataset;**
- **Flag for whether dataset is in agency search list;**
- **Name of model(s) that found the dataset;**
- **Threshold chosen for inclusion;**
- **Count of number of mentions;**
- **Count of number of unique research outputs in which that dataset has been found;**
- **Count of number of publications for each research topic.**

For unknown datasets with frequency greater than 5:

- **Predicted name of unknown dataset;**
- **Count of number of mentions;**
- **Count of number of unique research outputs in which that dataset has been found;**

- **Count of number of research outputs for each research topic.**

IDIES displays this in admin dashboard and records milestone in databases. **ADMIN** inspects the result and approves continuation. [precise details TBD].

Step 5: Generation Publication Record Data for validation [Elsevier]

Elsevier takes the research outputs that are matched with one or more target datasets and produces the agreed metadata for those research output records (see Appedix B for metadata that is available for individual research outputs). The data for individual research records is for those that contain the target datasets or their aliases only.

Whilst research output metadata is produced only for the research output that contain a target dataset or alias, it is possible that a research output will contain other datasets in addition to the target ones. In those circumstances, the snippet associated with those additional datasets will be provided.

Step 6: Ingestion in database [IDIES]

- **IDIES** retrieves data from S3 storage on SciServer;
- JSON files read and ingested into staging database;
- Data from staging tables transformed into core database;
- Admin dashboard shows statistsics such as:
  - total number publications found;
  - number of research outputs for each dataset;
  - number of topics for each dataset show topics sorted in descending frequency;
  - number of authors for each dataset;
  - number of journals for each dataset.

Step 7: Validation [IDIES]

ADMIN configures validation:

- defines users (email, password);
- assigns users as **REVIEWER** for this project;
- sets number of snippets in a batch;
- sets fraction of snippets whith multiple reviews.

REVIEWER validates snippets assigned to them.

ADMIN inspects progress of validation in admin dashboard.

V2 ADMIN provides map of EL research topics to AGENCY topic of interest.

Step 8: Finalization [IDIES]

ADMIN decides validation is complete.

Validated data is sent to S3 bucket.

- One CSV file per table.
- Only accepted dyads, no snippet data

TACC retrieves it and loads it in database underlying the API.

Elsevier extracts information of use for ML tuning.

APPENDIX D: SHOW US THE DATA WORKSHOP RESULTS

Chief Data Officers and Evidence Act Officials

On September 21, 2021, the Coleridge Initiative convened a panel of Evidence Act Officials and Chief Data Officers to request their input about the results of the "Show US the Data" competition before the October 20th workshop. The goal of the CDO expert pre-session was to develop a point of view from representative CDOs regarding the applicability of the learnings and tools developed in the "Show US the Data" competition. Since the competition focused on uses of data sets in research, the outcomes were most immediately appliable to agencies with scientific mission components. CDOs from agencies for which discovery activities occurred in the competition were invited to review results in one-on-one sessions and then to attend this panel discussion. The Agencies represented in the discussion were Commerce (NOAA), NSF, USDA, Transportation. Given that the breadth of data work in an Agency may cross many mission teams, some agencies had multiple team members participate in the discussion session.

Specific questions were posed to identify ways that the approach and prototype algorithms might be used to support agency mission activities both near-term and strategically, including (1) how the capabilities might be used; (2) opportunities for near term use in the agencies; (3) potential obstacles to use; (4) key points of engagement; (5) proposed next steps.

The type of tools developed in the competition might support emerging research themes, connect researchers to previously undiscovered datasets for stimulating new discovery, and providing evidence of citizen benefits. As such, they are more useful for prioritizing resources and work efforts for making public data available for research and public uses than as simply a pathway to achieve compliance with the Open Public Electronic Necessary Government Data Act and other mandates. They can drive broader visibility and transparency about datasets and their uses both within and outside the agencies. Most impactful would be creating communities around connecting and creating meaningful exchanges between the users of the data and those producing and maintaining the data.

One important barrier to use is the agencies' lack of current workforce skills for developing and using these types of technical tools. Also discussed were competing priorities for resources within agencies and overall priorities of Agency mission activities.

Building greater visibility and engagement would require a significantly expanded awareness outreach effort that could include Boards, special purpose groups, councils and civic tech organizations.

The discussions about near-term use and next steps coalesced into a common point – identification of specific use cases within Federal agencies to sponsor application of the approach and tools followed by analysis and capture of learnings from each step along the process-priority setting, workforce, barriers and engagement model.

It was suggested by the participants that the October session include some dialogue about potential use cases so that there might be collective sponsorship and support for the next steps.

Academic Researchers

The Coleridge Initiative convened a panel of experts representing the perspective of researchers to get their input about the results of the "Show US the Data" challenge before the October 20 conference. The goal of the panel was to gain understanding about a new machine learning approach to identifying public uses of agency data, identify strengths and weaknesses of the approach, discuss how researchers would draw on this usage information captured by live data streams, and suggest ways to incorporate feedback from the public on both the usage documentation and on the data sets. The panel discussion will be summarized and incorporated in the October conference.

The participants were asked a series of structured questions including: (1) how they might use the tools to advance their research; (2) how the tools might advance the work of junior researchers; (3) how the tools might inspire researchers to do their work differently; and (4) how might the researcher community become engaged in this effort? Below are some key highlights summarized. Detailed minutes are attached.

- Providing the right incentives for researchers facilitates success and encourages use and feedback to improve the system. The tools can assure that the burden is not all on the researcher to provide their publication data. Rather, a positive feedback loop could be created by researchers having their citations and publications included, getting people to advertise their work, giving seminars, and sharing their data and best practices in terms of citing data, so that their work can get acknowledged. This also could lead to improvements such as a uniform citation for a dataset.
- The tools allow researchers to make connections between what datasets are being used and for what purposes—allowing researchers to build on what's already been done. Making connections also highlights which datasets may be underused. The tools can also foster partnerships between academic researchers and government agencies that have data the researchers are using. Those two-way relationships can also help improve the data sets' accuracy and usability.

Publishers

The Publisher workshop was held in conjunction with three other workshops (Chief Data Officers, Researchers, and Academic Institutions) to answer questions and gather input to feed into the Coleridge Initiative "Show US the Data Conference" on October 20, 2021. Structured questions were asked to get feedback on what the academic institution stakeholder community thinks about Machine Learning/Natural Language Processing.

The workshop participants were asked structured questions to get feedback on what the publisher stakeholder community thinks about the potential of the Rich Text Content project and its machine learning and natural language processing components. (1) Concerns about the Machine Learning / Natural Language Processing (ML/NLP) approach to capturing data use; (2) Additional functionality that would be useful; (3) the value proposition for publishers to participate; (4) How publishers could participate; and (5) where should the application reside and be managed?

The participants raised several points including:

- This initiative needs to be a sustainable infrastructure where there is funding for work that is produced, and there is value in producing a high-quality curated corpus. There should be transparency in any pricing model. The small to medium publishers have valuable content and contributions but have a lower level of sophistication which may impact the rate of adoption.
- There should be a central place, such as data.gov, where this information can be accessed. In addition, a publisher dashboard maintained for smaller publishers could be very helpful so that publishers could also see how data are being used, citations, and new services that publishers could provide.
- One of the biggest challenges with reusing and understanding the ongoing value of datasets is how much metadata is there and how much context there is around that data. Researchers are also funded in a way that they don't have access to those government repositories and are left with fewer choices to put their data so they end up in general repositories like Figshare, Dryad, etc. General repositories aren't very helpful for building on research unless they are able to pull in the required metadata. There is a need for greater incentive for authors to comply with open data policies. If publishers make this more findable and prominent and enable credit as a first class object: incentives, quality, services, and compliance will increase.

Other discussion points raised included:

- *Value Proposition*: many publishers are investigating services that they might provide in relation to identification and analysis in getting data. Is this a free substitute for something that they would like to provide a service as part of a publisher's offerings? What is the value proposition for publishers?
- *Bias*: Having machine learning drawing conclusions about how data are being used may not lead to the most accurate insights. How can human interaction be added into the model to improve the results and the accuracy will continue to grow.
- *Relative importance of two main use cases*: (1) a compliance driven use case – for agencies to show that they are tracking reuse per the mandate; and (2) providing a means to discover data. To what extent has the relative importance of these use cases been established with users?
- *Risk of using NLP to capture data*: in making our entire Full text XML corpus available to do the work, how to ensure the content was only used for this purpose, by a controlled group, and deleted afterwards. This doesn't relate to concerns about the job itself (publishers do make content available to third parties for indexing, abstracting, etc).
- A link back to the publisher is critical, ultimately building an informal citation network. The community develop different visualizations to suit their needs?
- Publishers can participate in multiple ways, including allowing indexing services to use their content for this purpose; or running the ML/NLP algorithms internally on their content. Harvesting is currently allowed by some publishers.
- A public/private partnership that would be friendly to international users and be resilient to

US administrative government change could be considered to run this function, with central access available.

- A broad general value proposition is enhancing your value (both quantitatively and qualitatively) to the community you are trying to serve. There are also non-financial benefits for publishers, such as increased usage and citations. Publishers want to comply with funder goals, long term solution is more around formal citation (as mentioned earlier), and Value proposition on "win-win" content/data discovery, links between the two (more consumers of government data and more consumers of published articles) the technical and business challenges in creating a long-term solution? Could a combined effort be established?
- There is a need for equity across publishers and potentially there was additional enabling needed for smaller publishers. Most are doing things in a different way, so thinking in terms of using a broker where there is a degree of standardization may be a good idea.

The participants agreed that it may make sense to start off with a pilot on one or two specific topics.

- The tools foster community and mentorship, helping junior researchers use data to knit together people and research to impact their work. Junior researchers could discover new data sets and ways to use existing data sets and gain visibility if their research is represented in the database.
- An interactive partnership approach between agencies and the research community can help agencies prioritize by seeing how data are being used by others. Agencies can use the buy-in of researchers—documented as high use of certain datasets—to demonstrate its importance to Congress, call attention to underutilized data, and make investments in data improvement.

Several ideas were put forth to encourage researcher community engagement:

- Researchers could be incentivized by providing curation tools—for example, finding related datasets by joining datasets and cleaning up the data. Agencies could provide links to tools for cleaning and linking the data. Some researchers may not know how much data are available to them from agencies. Some younger researchers find this out only by asking more senior researchers.
- Access to grey literature (research that has not been published in a peer reviewed journal but are available in libraries of universities and elsewhere) could be incredibly valuable, creating communities around working papers and even avoiding publication bias.
- The tools offer further opportunity for development, such as information regarding authors (name, email addresses, etc.) could be harvested or a Citation Index drawn on for collaborators, allowing authors and other researchers in the field to build up a network to share information about certain metadata and otherwise clarify uncertainties, fill gaps, and improve overall use. Researcher could automatically share information about the quality of a dataset.

The panel identified the key next actions as follows:

- The project should stay focused on the value add and allow for exciting developments.

- Researchers should be able to see how easy the tool is to use and immediately see the value. Engaging high profile users—research "influencers"—could also be a great way to set a trend among others.

Academic Institutions

The Academic Institutions workshop was held in conjunction with three other workshops (Chief Data Officers, Researchers, and Publishers) to answer questions and gather input to feed into the Coleridge Initiative Show US the Data Conference on October 20, 2021. Structured questions were asked to get feedback on what the academic institution stakeholder community thinks about Machine Learning/Natural Language Processing.

The participants discussed several issues and brought up the key points below:

- Several benefits for researchers at institutions included improved discovery of what data exist and are available, better access to data, and opportunities for collaboration, especially across disciplines. More use of the data would also create motivation to improve the metadata, e.g., developing and conforming to metadata and citation standards and making sure data are complete. This would also help improve existing governance structures and help integration across existing infrastructures.
- Institutions want to understand usage and improve discovery and access from their data repositories. Institutions also use a lot of state and other data, so there could be wider applications beyond federal data. The application could also help identify gaps in which data were being underutilized. In addition, preservation policies for data could use data to support decisions, e.g., a librarian to check after a period of time to see if data has been used, and if no one has used the data, could archive it or stop maintaining it. The cumulative costs of maintaining repositories are going to be important and will be impacted by the use case of determining which datasets should be kept for what period of time.
- Some land grant universities have close relationships with federal agencies such as USDA. It could be helpful to consider pilot projects that build on these relationships.

Miscellaneous

Other discussion points included:

Concerns. Participants expressed a desire for data beyond those in scientific publications ("gray" literature, other media), ensuring the accuracy of data included in the dashboards, and establishing a mechanism for feedback on what agencies are doing with public comments and suggestions on improving the data. Privacy concerns related to the ability of competing institutions and researchers to view the dataset details that an institution is using, particularly prepublication. Questions arose on who would run such a service and be responsible for protecting privacy, uncovering potential bias, and usability.

Participation and access. A central host was generally favored but institutions also wanted to be able to host specific search and display capabilities particularly if "gray "literature that is held by an institutional library could be included. Possibilities include institution repositories or

research information management systems or library discovery environments. For example, University of Michigan has already invested into building out the crosswalks between the institution repository and the research information management system. The usage data should also be available to view at the site where the agency is providing the data.

Value. Data citations could lead to tenure or other salutary job impacts for researchers. In addition, Consortium approaches often work if incentives are created to benefit individual institutions and the group as a whole. Helping to rationalize the current system would also provide value as there are competing tools, and it's unclear which data are where, in what format, and in what detail. If agencies would provide standard citation info for the data sets that would be helpful.

References

[1]: Lane, J., I. Mulvany, and P. Nathan, *Rich Search and Discovery for Research Datasets: Building the Next Generation of Scholarly Infrastructure*. 2020, London: Sage.

[2]: Lane, J., et al., *Data Inventories for the Modern Age? Using Data Science to Open Government Data*. Harvard Data Science Review, 2022.

[3]: Oliveira, A.S., et al., *Prospective scenarios: A literature review on the Scopus database*. Futures, 2018. **100**: p. 20-33.

[4]: Burnham, J.F., *Scopus database: a review*. Biomedical digital libraries, 2006. **3**(1): p. 1-8.

[5]: Aghaei Chadegani, A., et al., *A comparison between two main academic literature collections: Web of Science and Scopus databases*. Asian social science, 2013. **9**(5): p. 18-26.

[6]: Kaggle. Kaggle: Show US the Data. 2021 02/09/2022]; Available from: https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data

[7]: Ghani, R., *The Winning Methods, in Show US the Data*. 2021, Coleridge Initiative https://coleridgeinitiative.org/wp-content/uploads/2021/11/Coleridge-Conference-Deck-rayid-ghani.pdf: Virtual

[8]: The Coleridge Initiative, *Show US the Data Final Report*, C. Initiative, Editor. 2021: Coleridge Initiative.

[9]: Szalay, A.S., et al., *Designing and mining multi-terabyte astronomy archives: The Sloan Digital Sky Survey*. ACM SIGMOD Record, 2000. **29**(2): p. 451-462.

[10]: Szalay, A.S., *From skyserver to sciserver*. The ANNALS of the American Academy of Political and Social Science, 2018. **675**(1): p. 202-220.

[11]: Taghizadeh-Popp, M., et al., *SciServer: A science platform for astronomy and beyond*. Astronomy and Computing, 2020. **33**: p. 100412.