







Show US the Data Conference
October 20, 2021
Final Report

COLERIDGE INITIATIVE

Show US the Data Conference Report

Contents

Summary 2
Conference details 3
Appendix 1: Researcher Panel Executive Summary 6
Appendix 2: Academic Institutions Panel Executive Summary 8
Appendix 3: Publishers' panel Executive Summary 10
Appendix 4: CDO Panel Executive Summary 12

Summary

The **Show US the Data** virtual conference on October 20, 2021 launched the next stage of a new approach to enable federal agencies and their stakeholders to use new tools and technologies to show how their data are used. The conference was extremely successful. First it showed that automated search and discovery tools were possible to develop. Second, it made it clear that those tools could be used to find and close some of the gap between what agencies know about how their data are being used and how much data use is actually taking place. Third, it served to identify a broad community of stakeholders ready to work with agencies to close the gap even further. Finally, a number of agencies agreed that there was a great deal of potential to move forward with use cases that would highlight the value of their data in supporting evidence based policy-making.

The proposed work is not occurring in a void. There is massive interest in increasing the use of data. Title II of the Evidence Act mandates that federal agencies produce data inventories about the use of their data and requires them to work with the public to do so. The Federal CDO Council has issued a Request for Information to help them in the activity.

The new approach can help agencies answer critical questions. How can federal agencies show how their data are used? How can they make decisions about what data to invest in? How can researchers, academic institutions, and publishers help build data and evidence to better inform policy? How can they both find the gap in knowledge about how federal data are being used and close the gap?

The conference, which was organized by former Federal CIO, Suzette Kent, former US Chief Statistician, Nancy Potok, and Julia Lane, all of the Coleridge Initiative, and Howard Ratner of CHORUS brought together a wide diversity of public data users, data producers, researchers, academic institutions, publishers and federal agencies. A recording of the conference is now available.

The core of the approach is to automate the search for and discovery of the use of federal datasets. Agencies can then respond to the Evidence Act, researchers and academic institutions can receive credit for promoting and supporting data use, and publishers can foster the reproducibility and replicability of empirical research. The conference highlighted new tools that were developed through a Kaggle competition, which can be expanded to include the usage of data in government reports, the press, and by the public at large.

Conference details

Suzette Kent, former Federal CIO, opened the conference with welcoming remarks and with a rousing call to action. She noted the importance of highlighting collaborations and use of technologies to break down barriers and using federal data for public good.

Former Speaker, Paul Ryan, provided the legislative context and highlight the importance of data and evidence in policy making in his keynote speech. He spoke to the importance of the Evidence Act Title 2 – Open Government Data Act -- to facilitate collaboration with non-government entities, researchers, and the public for the purpose of understanding how users value and use government data. He is excited about seeing the Evidence Act in action and identifying next steps so we can effectively put the promise of using data to inform decision-making in the public sphere into practice.

Julia Lane, co-Founder of the Coleridge Initiative, then provided background on the motivation for and structure of the Kaggle competition in her remarks. She pointed out that the motivation for the Kaggle competition was to identify the gap in knowledge about data and then close that gap to inform better decision-making. The Kaggle competition showed that it was feasible to develop Machine Learning and Natural Language Processing tools to find datasets in scientific text: more than 1,600 teams from around the world entered the competition.

Rayid Ghani, Professor at Carnegie Mellon University, provided an overview of the winning algorithms. He noted that the task was not easily tackled applying off the shelf Machine Learning/Artificial Intelligence methods. It required understanding the problem, the use case (and the metric of interest), looking at the data, and creating custom solutions. He pointed out that the winning solutions ranged from manual, to human combined with some Machine Learning, to humans generating variations of data and heuristics to make Machine Learning work for them. He pointed out that efforts like this need collaborations across disciplines, exposing computer scientists and social scientists to real-world problems and incentivizing them to create solutions.

Representatives from three agencies (US Department of Agriculture Economic Research Service, the Department of Commerce National Oceanic and Atmospheric Administration, and the National Science Foundation/National Center for Science and Engineering Statistics) then demonstrated how the winning algorithm could be used. Each agency took a somewhat different approach, and Spiro Stefanou (Administrator of the Economic Research Service, USDA) , Thomas Beach (CDO, USDOC) , Dorothy Aronson (CIO, NSF) and Vipin Arora (Deputy Division. Director, NCSES, NSF) presented their agencies preliminary visualizations utilizing a database of results created by using methods from the Kaggle competition winners. Visualization tools such as word clouds (for USDA) , Tableau (for USDA and NOAA) and Qlik (for NSF) were used to produce dashboards that display a range of outputs, allowing individuals to

track how agency data are being used and by whom. The interactive visualizations included word clouds and dashboards that greatly expanded the knowledge of the agencies. Importantly, they made it clear that the tools will help the agencies meet the mandates of Title 2 of the Foundations of Evidence Based Policy Act, known as the Open Government Data Act, which requires agencies to determine how their data are being used by the public and to interact with the public to improve their data quality and accessibility.

A panel discussion of key stakeholders and members of the new data ecosystem was initiated by Tim Persons, the Chief Scientist of GAO. The panel reported the results from a series of workshops held in September 2021 by CHORUS and Coleridge with US federal agencies, researchers, academic institutions, and publishers, including summaries of stakeholder benefits, challenges, opportunities, and next steps.

Kaye Husbands Fealing (Dean, Ivan Allen College of Liberal Arts at Georgia Institute of Technology) reported on behalf of the research panel that the proposed approach could provide incentives for researchers to facilitate success, encourage use, and build on what's already been done. It is extremely important to make connections between what datasets are being used and for what purpose because it allows researchers to build on what's already been done. She argued it was important to build an interactive partnership between agencies and the research community (panel summary in Appendix 1).

Judith Russell (Dean, University Libraries at the University of Florida) reported on behalf of the academic institutions panel that the tools offer an opportunity to improve discovery of what data sets exist with enhanced information about and access to that data. She urged that the approach be expanded to find the availability of data beyond that included in scientific publications. She emphasized the importance of ensuring that the tools supported the accuracy of data, protection of privacy, and uncovering bias. The academic institutions recognized the value of a central host while recognizing the desire for agencies or academic institutions to host specific data and services. They also valued the potential for collaboration opportunities between academic institutions and federal agencies to use these tools (panel summary in Appendix 2).

Alix Vance (Chief Executive Officer, AIP Publishing) , noted on behalf of the publishers the importance of a sustainable infrastructure with a central location for access. She also noted both how important it is to link back to the article DOI, and that it will be challenging to understand just how much metadata are available. She suggested that there were multiple ways for publishers to participate, including establishing private/public partnerships (see Appendix 3).

Dan Morgan, Chief Data Officer, US Department of Transportation, noted that agencies would need to focus on three areas.

The first of these is establishing *impact*: The most compelling value that these types of tools might offer for agencies is prioritizing resources and work efforts for making public data available for research and public uses. In addition, these types of tools offer a mechanism for connecting communities and creating meaningful exchanges between the users of the data and those producing and maintaining the data. The second is addressing *challenges*: One important challenge to the introduction of new approaches and tools is competing resource priorities at Agencies. In addition, there is limited expertise in many Agencies in use and maintenance of these types of technical tools. The third is coming up with *compelling next steps.* 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 (see Appendix 4).

A key takeaway from these four stakeholder groups was the importance of identifying next steps so we can effectively put the promise of using data to inform decision-making in the public sphere into practice using machine learning and natural language processing. The researchers, institutions, publishers, and agencies all agreed that a logical next step would be to move forward with pilot projects to build momentum within the data ecosystem by providing the tools to link the participants, fostering collaborations that will identify the gaps in the data and fill the gaps towards the end of creating better evidence to inform agency operations and policies.

The conference was closed out by Nancy Potok (former Chief Statistician of the US) who outlined potential next steps for pilot projects involving some tangible advances for the presenting agencies and stakeholders over the next several months that have immediate value. The pilot projects would show how important datasets have been used in particular areas by finding and connecting data producers and data users, measuring how data are being used for evidence, and sharing knowledge about innovative tools, metadata, APIs, and dashboards. The results of this information could be used immediately by the Federal CDO Council, agencies and stakeholders and serve as a basis for further adoption of the tools and continuing to improve their usefulness through collaborative efforts. Howard Ratner, CEO of CHORUS, who had moderated the entire conference, then concluded the event.

To stay informed and involved with the newly forming data ecosystem, please register here.

Appendix 1: Researcher Panel Executive Summary

Tuesday, September 21, 2021

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? The following provides a summary of some key highlights.

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.

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, namely:

- 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.

Appendix 2: Academic Institutions Panel Executive Summary

September 21, 2021

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.

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.

Appendix 3: Publishers' panel Executive Summary

September 2, 2021

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. 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?

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.

Appendix 4: CDO Panel Executive Summary

Tuesday, September 21, 2021

On September 21, 2021, the Coleridge Initiative convened a panel of 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 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; and (5) proposed next steps.

A few overarching themes emerged:

- These tools can 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 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' cultural resistance to change. Other barriers include lack of current workforce skills and competing priorities for resources within agencies.
- 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.