Rich Context Introductory Chapter

Ian Mulvany, Paco Nathan, Sophie Rand, Julia Lane

# Introduction

Science is at a crossroads. The enormous growth of access to data coupled with rapid technological progress, has created opportunities to conduct empirical research at a scale that would have been almost unimaginable a generation or two ago. Researchers can now rapidly acquire and develop massive, rich datasets; routinely fit complex statistical models; and conduct their science in increasingly fine-grained ways. Yet there is no automated way to search for and discover what datasets are used in empirical research, leading to fundamental irreproducibility of empirical science and threatening its legitimacy and utility(*1*, *2*). There is an enormous interest to change the current manual and ad-hoc system, and incentives are increasingly aligned: while only a fraction of datasets are identified in scientific research, those publications that do cite data are cited up to 25% more than those that do not(*3*).

Vannevar Bush foreshadowed the issue more than 60 years ago:

“There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers—conclusions which he cannot find time to grasp, much less to remember, as they appear. … Mendel’s concept of the laws of genetics was lost to the world for a generation because his publication did not reach the few who were capable of grasping and extending it; and this sort of catastrophe is undoubtedly being repeated all about us, as truly significant attainments become lost in the mass of the inconsequential”(*11*).

We can do better – and we now have the opportunity to do so.

The core problem that needs to be addressed is automating the search for and discovery of datasets used in empirical data – building an Amazon.com for data. The vast majority of scientific data and outputs cannot be easily discovered by other researchers even when nominally deposited in the public domain. Faced with a never-ending stream of new findings and datasets generated using different code and analytical techniques, researchers cannot readily determine who has worked in an area before, what methods were used, what was produced, and where those products can be found. Resolving such uncertainties consumes an enormous amount of time and energy for many social scientists. A new generation of automated search tools could help researchers discover how data are being used, in what research fields, with what methods, with what code and with what findings —often by passively capitalizing on the accumulated labor of one’s extended research community. And automation can be used to reward researchers who validate the results and contribute additional information about use, fields, methods, code, and findings.(*8*)

New advances in technology—and particularly, in automation—can now change the way in which social science, and hence other sciences, is done. The place to start is with the social sciences. The great challenges of our time are human in nature - terrorism, climate change, the use of natural resources, and the nature of work - and require robust science to understand the sources and consequences. The lack of reproducibility and replicability evident in many fields(*1*, *4*–*7*) is even more acute in the study of human behavior both because of the difficulty of sharing confidential data and because of the lack of scientific infrastructure. Social scientists have eagerly adopted new technologies in virtually every area of social science research—from literature searches to data storage to statistical analysis to dissemination of results(*8*). And, in the United States, the recent passage of the Foundations of Evidence-based Policymaking Act(*9*, *10*) and the focus on a Federal Data Strategy, mean that there is an important use case for showcasing the value of new approaches.

The knowing how it has been produced and used before: the required elements what do the data ***measure***, what ***research*** has been done by what ***researchers,*** with what ***code***, and with what ***results***. Acquiring that knowledge has historically been manual and inadequate. The challenge is particularly acute in the case of confidential data on human subjects, since it is impossible to provide fully open access to the source files.

# How this book came to be

This book was born out of a need to solve a very concrete problem. In 2016, the US Congress passed the Commission on Evidence-based Policymaking Act to make a set of recommendations on how to better use data for decision-making. The US Census Bureau was charged with supporting the deliberations of the Commission and asked our team at New York University to build a secure access facility in which data from multiple agencies could be securely hosted.

After we built the facility, and had dozens of users, we realized that putting data in one place, while necessary, was not sufficient for good analytical work to be done. Every user who accessed the data wanted to know what other work had been done with the data, with what assumptions and what results. We were able to provide them with some information, but essentially the information was drawn from our own research experience and was certainly not representative of the entire field. The obvious solution was to see if computer scientists had the technological tools available to automate the discovery of research datasets and the associated research methods and fields in research publications. Our computer science colleagues assured us that, while the technology existed in principle, no single team was known for having developed a solution.

We decided to see what we could to advance the field, and approached Schmidt Sciences, the Alfred P. Sloan Foundation and the Overdeck Family Foundation for support. As part of that support, we ran the competition with the results described in this book. We challenged participants to combine machine learning and natural language processing methods to identify the datasets used in a corpus of social science publications and infer both the scientific methods and fields used in the analysis and the research fields.

The core of the book describes both how the competition was set up, as well as the results achieved by different competing teams. However, as is always the case with exciting research agendas, the competition helped us identify five major scientific challenges that need to be addressed: (i) document corpus development, (ii) ontology development for dataset entity classification, (iii) natural language processing and machine learning models for dataset entity extraction, (iv) graph models for improving search and discovery, and (v) the use of the results to engage the community to both validate the model results, retrain the model and to contribute code and knowledge. So the other chapters in the book provide an overview of what could be done with more resources and talent devoted to this interesting question. The next section provides a more detailed overview of the contribution of each chapter.

# Book overview

Section 1 provides an overview of the motivation and approach. Section 2 describes new approaches to develop corpora and ontologies. Section 3 describes the competition results in terms of model development. Section 4 provides a forward looking agenda.

Section 1: Motivation and approach

In Chapter 2, “ Where’s Waldo: Conceptual issues when characterizing data in empirical research,” researchers from the Research Data and Service Center at the Deutsche Bundesbank show us why better metadata for social science data enables discovery of datasets and research, in ways that surpass what traditional metadata from data producers can support. They present a new modus operandi in the service delivery model of research data facilities, based on the premise that datasets have a measurable value that can be deduced from the relationships between datasets and publications, and the people who author, do research on, and consume them - that is, Rich Context around datasets.

They argue that a major advantage of rich context is that it closes the loop on metadata is closed: a loop initiated by the metadata from the data producer side, is closed by metadata from the data usage side. The authors elucidate why such rich data from the *usage* perspective is needed to deliver codified knowledge to the research community to guide literature review and new research; without understanding the linkage between datasets and outcomes, we are disabled in shaping new, impactful research.

The authors identify two primary reasons for this need: first, that metadata on the datasets from the data users perspective helps the data creator to improve upon the quality of the data itself, improving dataset owners’ service delivery (e.g. bundesbank as a service provider, the service being data provision, consulting on dataset usage, creation of new data products, etc); and second, that metadata on the usage of datasets in publications helps us measure impact of datasets in their ability to drive policy-making. With this closed loop, the scope of researchers’ discovery is broadened to include not only literature and datasets, but the interplay between those two - that is, how datasets have been used by whom and how.   The authors discuss a tangible outcome of measuring dataset value - a dataset recommendation system, enabling expedient sharing of available datasets through the research community.

Chapter 3 outlines the operational approach that was taken to develop the [Rich Context Competition](https://coleridgeinitiative.org/richcontextcompetition). The goal of the competition, the results of which are described in Section 2, was to implement AI to automatically extract metadata from research - identifying datasets in publications, authors and experts, and methodologies used. As such, the competition was designed to engage practitioners in AI and NLP to develop models based on a corpus developed at the Interuniversity Consortium of Political and Social Research. The competition attracted 20 teams from around the world and resulted in four finalists presenting their results at NYU on February 15, 2019 (see the [agenda and video here](https://coleridgeinitiative.org/richcontextcompetition/workshopagenda)).

The results of the competition provided metadata to describe links among datasets used in social science research. In other words, the outcome of the competition generated the basis for a moderately-sized knowledge graph. the [winning team](https://ocean.sagepub.com/blog/an-interview-with-the-allen-institute-for-artificial-intelligence) in the Rich Context Competition was from [Allen AI](https://allenai.org/) which is a leader in the field of using embedded models for natural language. Typical open source frameworks which are popular for deep learning research include [PyTorch](https://pytorch.org/) (from Facebook) and the more recent [Ray](https://ray.readthedocs.io/en/latest/distributed_training.html) (from UC Berkeley RISElab).

# Section 2:

A major challenge is developing a training corpus that sufficiently represents the population of all documents, and tagging the datasets in the corpus. It is essential to do this well if high quality models are to be developed. There is a literature outlining the issues with developing a "gold standard corpus" (GSC) of language around data being mentioned and used in analysis in academic publications, since creating one is time-consuming and expensive (*12*) In Chapter 4 “Standardized Metadata, Full Text and Training/Evaluation for Extraction Models”, Sebastian tk and Alex Wade describe the need for, and strategies for collecting, large sets of annotated full-text sources for use as training data for supervised learning models developed in the Rich Context Competition. Dataset Extraction, the NLP task at the core of the Rich Context Competition, relies on a high-quality set of full text sources with metadata annotations. Developing such a corpus must be done strategically, as full-text articles and their metadata are organized inconsistently across their sources. The corpora gathered for use as training data for the Competition required ad-hoc manual labor to compile. Here, authors describe the legal, technological and human considerations in creating a corpus. They dictate the scale of full-text data needed, and the impact that an interdisciplinary (e.g spanning multiple domains) corpus has on that scale. They suggest development of a corpus with open-access text resources, use of human-annotators for labeling of full-text, and attention to the mix of domains that may be in a corpus when developing models.

There is a separate challenge of developing a common understanding of what a dataset is. Developing standard ontologies is a fundamental scientific problem, and one that is often in the domain of libraries and information scientists. Although some measure of linguistic ambiguity is likely to be unavoidable in the social sciences given the complex subject matter, even modest ontologies that minimally control the vocabulary researchers use would have important benefits. In Chapter 5, “Metadata for Administrative and Social Science Data”, Robert B. Allen describes a framework for the application of metadata to datasets, details existing metadata schema, and gives an overview of the technology, infrastructure and human elements that need to be considered when designing a rich metadata schema for describing social science data.

Allen describes three types of metadata - structural, administrative and descriptive; and emphasizes the growth needed in descriptive metadata, which are characterized by semantic descriptions. Allen describes existing metadata schemas which accommodate domain-specific metadata schema, like the W3C DCAT, and the unique semantic challenges faced by social science as opposed to natural sciences - in particular that concepts - e.g. “family”, “crime” -  are less well-defined, and definitions change across sub-domains. He considers data repositories and describes the essential role of metadata in making such repositories searchable and therefore useful. He touches on several prominent data repositories in the social and natural sciences and describes their methods of gathering metadata and how the metadata supports services offered, like search, computing environments, preservation of data for archives, and logging of the history of a dataset and its provenance. Allen describes other challengings in creating and maintaining metadata, prompted by things like changes in technology that yield data streams, and changes in metadata standards. He discusses some of the technology underlying data repositories; in particular data cubes for data storage that facilitate data exploration and retrieval; containerization and cloud computing enabling sharing and reproducibility; and collection management systems which can provide metrics on usage, like number of downloads, maintenance of datasets, etc.

# Section 3:

Chapter 6, by the Allen AI team, describes their overarching approach. The team used a named entity recognition model to predict dataset mentions. For each mention, they generated a list of candidate datasets from the knowledge base. They also developed a rule based extraction system which searches for dataset mentions seen in the training set, adding the corresponding dataset IDs in the training set annotations as candidates. They then use a binary classiﬁer to

predict which of these candidates is a correct dataset extraction. While this approach was eventually the winning approach given the design of the corpus and the scoring mechanism, it suffers from being too fragile for general application, since it is necessarily corpus dependent. That team did not devote substantial time to identifying fields and methods.

Chapter 7, by the KAIST team, describes a very different approach. They generated their own questions about dataset names and use a machine learning technique to train the model for solving question answering task. In other words, questions suitable for finding dataset names such as “What is the dataset used in this paper?,” are generated and the question answering model is trained to find the answers to those questions from the papers. Furthermore, the resulting answers from the model are filtered by types of each word. For example, if an answer contains words with organization or agency types, then this answer is likely to include the actual dataset names. They also were quite innovative with identifying research fields, by using Wikipedia as the source, and methods by using machine learning techiques

Chapter 8, by the GESIS team, also used a Named Entity Recognition procedure. However, their design was module-based approach and they developed tools that can be used separately but also as parts of a data processing pipeline. For identifying research methods and fields, they exploited the Social Science Open Access Repository maintained at GESIS – Leibniz Institute for the Social Sciences. They also used the ACL Anthology Reference Corpus which is a corpus of scholarly publications about computational linguistics

Chapter 9, by the DICE team at Paderborn University, also used a Named Entity Recognition approach. They trained an Entity Extraction model based on Conditional Random Fields and combined it with the results from a Simple Dataset Mention Search to detect datasets in an article. For the identification of Fields and Methods, they essentially used search string to find embedded words

Chapter 10, by Singapore Management University, was an incomplete submission, with a very interesting approach. They used dataset detection followed by implicit entity linking approach to tackle dataset extraction task. They adopt weakly supervised classification for research methods and fields identification tasks utilizing SAGE Knowledge as an external source and as a proxy for weak labels.

# Section 4: Looking forward

In Chapter 11, researchers from Digital Science describe the role user engagement plays in creating rich context around datasets, which are take on properties of ‘first class research objects’ (like journal articles) in that they are published as distinct research outputs in their own right.  Authors lay out a set of challenges in the sharing of datasets and dissemination of dataset metadata, and articulate goals in creating infrastructure to answer these challenges.

As technology has yielded ever larger streams of datasets available for social science research, two critical, interrelated elements of infrastructure have not kept apace: information infrastructure, and cultural infrastructure.  Information infrastructure refers to content of interest to the rich context competition models - journal articles, datasets, and their metadata (including details on the data stewards, usage of the datasets in research, and code used to prepare and analyze datasets). Cultural infrastructure refers to the incentives and value propositions in place to encourage individual data stewards, data users and experts to share datasets and contribute metadata on datasets. Cultural infrastructure around datasets do not fit into the existent culture of research publications.

In venturing to build out information infrastructure around datasets, we must consider how concepts like versioning, reproducibility, and peer review carry over to datasets. Further, how do metadata carry over, when there is so much variability in what we mean when we say dataset? Incentives around data sharing, dataset curation, and metadata contribution are even slimmer than in publishing, where there exists a culture of “publish or perish.” This question must be resolved if we wish to enrich the context around datasets to make them more efficiently consumable.

The future agenda is described in the concluding chapter by Paco Nathan and Ian Mulvany

The first step is to create a corpus of research publications to be used for training data during the Rich Context Competition.

The next step will be a formal implementation of the knowledge graph, based primarily on extensions of open standards and use of open source software. That graph is represented as an extension of a DCAT-compliant data catalog. Immediate goals are to augment search and discovery in social science research, plus additional use cases that help improve the knowledge graph and augment research.

In the longer term, the process introduces human-in-the-loop AI into data curation, ultimately to reward researchers and data stewards whose work contributes additional information into the system. With this latter step, in the broader sense Rich Context helps establish a community focused on contributing code plus knowledge into the research process

# More resources

General competition information

The competition had two phases. In the first phase, participants were provided with labeled data, consisting of a corpus of 2,500 publications matched to the datasets cited within them. Participants could use this data to train and tune their algorithms. In the second phase, they were provided with a large corpus of unlabeled documents and asked to identify the datasets used in the documents in a test corpus, as well as the associated methods and research fields. The participants were scored on the accuracy of their techniques, the quality of their documentation and code, and the efficiency of the algorithm – and also on their ability to find methods and research fields in the associated passage retrieval.

The timeline was as follows:

* **September 30 2018:** Participants submit a letter of intent (see [How to Participate](https://coleridgeinitiative.org/richcontextcompetition#howtoparticipate))
* **October 15 2018:** Participants notified and Phase 1 data provided (see [First Phase Participation](https://coleridgeinitiative.org/richcontextcompetition#phase1participation))
* **November 15 2018:** Preliminary algorithm submitted (see [Program Requirements](https://coleridgeinitiative.org/richcontextcompetition#programreqs))
* **December 1 2018:** 15 finalists selected (see [First Phase Evaluation](https://coleridgeinitiative.org/richcontextcompetition#phase1evaluation)) and Phase 2 data provided (see [Second Phase Participation](https://coleridgeinitiative.org/richcontextcompetition#phase2participation))
* **January 15, 2019:** The algorithms of up to 6 teams are selected for final submission (see [Second Phase Evaluation](https://coleridgeinitiative.org/richcontextcompetition#phase2evaluation))
* **February 15 2019:** Workshop is held in New York for final presentation and selection of winning algorithms (see [Second Phase Evaluation](https://coleridgeinitiative.org/richcontextcompetition#phase2evaluation))

All the information provided to participants was available here

[https://github.com/Coleridge-Initiative/rich-context-competition](https://urldefense.proofpoint.com/v2/url?u=https-3A__github.com_Coleridge-2DInitiative_rich-2Dcontext-2Dcompetition&d=DwMFaQ&c=slrrB7dE8n7gBJbeO0g-IQ&r=omwcNBUqPba9pikmkXZXk2bFQ7zxZPhI5OH9dd8lFDA&m=jJJRJpvbdwLAeHNwur9CtaqPIY6UXS4q64avAMUSVV0&s=abG_3lYZ3eu8BWs6kkau2rcXOIwLyiymwo0uj6vwGt0&e=)

# References

1. J. P. A. Ioannidis, Why Most Published Research Findings Are False. *PLoS Med*. **2**, e124 (2005).

2. M. R. Munafò *et al.*, A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 21 (2017).

3. G. Colavizza, I. Hrynaszkiewicz, I. Staden, K. Whitaker, B. McGillivray, The citation advantage of linking publications to research data (2019), (available at https://arxiv.org/pdf/1907.02565.pdf).

4. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637 (2018).

5. A. Dafoe, Science deserves better: the imperative to share complete replication files. *PS Polit. Sci. Polit.* **47**, 60–66 (2014).

6. N. Young, J. Ioannidis, O. Al-Ubaydli, Why Current Publication Practices May Distort Science. *PLoS Med* (2008).

7. G. Christensen, E. Miguel, Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.* **56**, 920–980 (2018).

8. T. Yarkoni *et al.*, “Enhancing and accelerating social science via automation: Challenges and Opportunities” (2019), , doi:10.31235/osf.io/vncwe.

9. N. Hart, T. Shaw, Congress Provides New Foundation for Evidence-Based Policymaking (2018), (available at https://bipartisanpolicy.org/blog/congress-provides-new-foundation-for-evidence-based-policymaking/).

10. Office of Management and Budget, Federal Data Strategy (2019), (available at https://strategy.data.gov).

11. V. Bush, *Science, the endless frontier: A report to the President* (US Govt. print. off., 1945).

12. L. Wissler, M. Almashraee, D. M. Díaz, A. Paschke, in *IEEE GSC* (2014).