Contributor Bios

Karam Abdulahhad is a postdoctoral at GESIS - Leibniz Institute for the Social Sciences in Germany. He is engaged in the ExploreData project to build an advanced search engine for social science data. He has a Ph.D. degree in computer sciences from Grenoble-Alpes University in France, where he tackled the problem of term-mismatch. He proposed a new IR model by adapting an idea from physics. His research interests include IR theory, logical/conceptual/semantic IR, machine learning, and text mining. Recently, he is studying the profitability of the modern embedding technics in IR. He taught in several universities and developed several tools.

Palakorn Achananuparp is a senior research scientist at Living Analytics Research Centre (LARC), Singapore Management University. He is interested in developing and applying machine learning, natural language processing, and crowdsourcing techniques to solve problems in a variety of domains, including online social networks, politics, and public health.

Daniel Acuna is an Assistant Professor in the School of Information Studies at Syracuse University, Syracuse, NY. He runs the Science of Science and Computational Discovery Lab, supported by grants from NSF, DDHS, and DARPA and featured in Nature Podcast, The Chronicle of Higher Education, NPR, and the Scientist. The goal of his current research is to predict future academic success and remove potential biases that scientists and funding agencies commit during peer review. He has created tools to improve literature search, peer review, and detect scientific fraud.

Bob Allen is developing a model-oriented approach to information organization. His previous work ranged from recommender systems to neural networks. Bob studied at Reed College and UCSD. He joined the Research organization at Bell Laboratories. He then joined to the Bellcore Applied Research group in information science and digital libraries. He was the Editor-in-Chief of the ACM Transactions on Information Systems and later Chair of the ACM Publications Board. Since 1998 he has been a faculty member at number of universities around the world such as Maryland, Drexel, Victoria (NZ), Tsukuba (Japan), and Yonsei (Korea).

Waleed Ammar is a senior research scientist at Google, where he works on NLP-related problems in biomedical and clinical applications. Before joining Google, Waleed was a research scientist at the Allen Institute for Artificial Intelligence where he led the Semantic Scholar research team. In 2016, he received a Ph.D. degree in artificial intelligence from Carnegie Mellon University. Before pursuing the Ph.D., Waleed was a software developer at Microsoft Research, web developer at eSpace Technologies, and teaching assistant at Alexandria University.

Christine Betts is a software engineer working on human computation at Google AI. She graduated with honors in computer science from the University of Washington. While there, she was an intern at The Allen Institute for AI, and before that at Facebook and Google.

Katarina Boland is research associate in the department Knowledge Technologies for the Social Sciences at GESIS - Leibniz Institute for the Social Sciences. She joined GESIS in August 2011 after earning her Magistra Artium degree in Computational Linguistics, Computer Science and Psychology at Heidelberg University. Katarina has been part of the DFG projects InFoLiS I and

InFoLiS II which addressed the automatic linking of research data and scientific publications. Katarina’s main research interests lie in the field of Natural Language Processing and Text Mining. Currently, she is primarily involved in research on Information Extraction, NLP & Journalism and automatic fact-checking.

Minh-Son Cao is a Master student in School of Computing at KAIST, under the supervision of Professor Sung-Hyon Myaeng at Information Retrieval and Natural Language Processing Laboratory. Previously, he received his Bachelor Degree from the University of Engineering and Technology, Vietnam National University (VNU-UET) in June 2017. He was a member of Data Mining and Knowledge Technology Laboratory from August 2015 to June 2017, under the supervision of Associate Professor Xuan-Hieu Phan. His research focuses on the application of Deep Learning in Natural Language Processing, mainly on embedding problems.

Stefan Dietze is full professor for Data & Knowledge Engineering at the Institute for Computer Science at Heinrich-Heine-University Düsseldorf, Scientific Director of the department Knowledge Technologies for the Social Sciences at GESIS – Leibniz Institute for the Social Sciences and affiliated member at the L3S Research Center of the Leibniz University Hanover, Germany. His research interests are at the intersection of information retrieval, semantic technologies and artificial intelligence, and in particular, the extraction, fusion and search of knowledge and data on the Web. Stefan’s work has been published at major scientific venues, such as WWW/The Web Conf, SIGIR, CHI or ISWC, where he also frequently serves as PC and/or OC member.

Dimitar Dimitrov is a PostDoc at GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany. He obtained a Ph.D. from the University of Koblenz-Landau, Koblenz, Germany. Before that, he studied Software Engineering at the University of Applied Sciences Konstanz, Konstanz, Germany, where he also obtained his master’s degree in Computer Science. At GESIS, Dimitar Dimitrov is working on the da|ra project aimed to deliver the software infrastructure for assigning DOI names to social and economic datasets. His research focuses on applying statistical and machine learning techniques to study user behavior in web-based systems.

Behnam Ghavimi is a research fellow in WTS department at GESIS – Leibniz Institute for the Social Sciences. He graduated from the University of Bonn with a Master’s degree in Computer Science. His master thesis was about detecting dataset references in texts under the supervision of Prof. Sören Auer and Dr. Philipp Mayr. Since September 2016, he was involved in different projects focused on NLP (text analysis and text mining) and recommender systems. One of his projects was the EXCITE project - jointly run by WeST at the University of Koblenz-Landau and GESIS to extract citations from publications and make more citation data openly available.

Andrew Gordon is Senior Data Engineer at Columbia University Information Technology. Previously, Andrew was Research Information Scientist with the Coleridge Initiative at New York University. There, Andrew served as an information specialist, programmer, and ETL engineer supporting the full research and administrative data lifecycle for ingest, curation, facilitating data discovery, and providing access to sensitive, administrative data for academics and policy analysts. Andrew has a Master of Science degree in Information from the University of Michigan School of Information and a Bachelor of Arts degree in Cultural Anthropology from the University of Michigan.

Suchin Gururangan is a Predoctoral Young Investigator at the Allen Institute for AI (AI2). His research interests involve model evaluation and robustness in NLP, especially under low-resource settings and distant domains. Before joining AI2, Suchin was a master's student in NLP at the University of Washington, and before graduate school, Suchin was a data scientist at various companies in Boston in Seattle.

Giwon Hong is a master's student in the School of Computing at KAIST and research assistant in IR&NLP Lab at KAIST. He graduated from Sungkyunkwan University with a degree in Computer Science in February 2018. His research lays in the area of Natural Language Processing, specifically in Question Answering and Relation Extraction

Rricha Jalota is a developer in the Computer Science department of Paderborn University. She works in the areas of data access and knowledge extraction. Her interests lie in the application of Machine Learning/Deep Learning approaches to solve NLP problems in the domain of Question Answering, Conversational AI and Information Retrieval.

Daniel King is a Predoctoral Young Investigator on the Semantic Scholar team at The Allen Institute for AI. He received his B.S. in Computer Science from Harvey Mudd College in May 2018. His research interests are generally in Natural Language Processing and using AI techniques to make useful tools. Outside of research and software engineering, he enjoys playing soccer, bughouse chess, and hiking.

Sebastian Kohlmeier is the Sr. Manager of Program Management and Business Operations at the Allen Institute for AI where he leads program management for applied research, business intelligence and data science and partner development. Prior to joining the Allen Institute for AI, Sebastian worked as a Technical Program Manager and Engineering Manager in a variety of roles at Amazon and Microsoft. Sebastian graduated with honors from Western Washington University in 2007.

Philips Kokoh Prasetyo
Philips Kokoh Prasetyo is a principal research engineer at the Living Analytics Research Centre (LARC) in the Singapore Management University. He enjoys analyzing data from many different perspectives, and his current interests include machine learning, natural language processing, text mining, and deep learning. He received Master degree from National Cheng Kung University in Taiwan, and Bachelor degree from Sekolah Tinggi Teknik Surabaya in Indonesia. He received several awards including ACLCLP thesis award in 2009, and DPU scholarship from 2007 to 2009.

Julia Lane is a Professor at the NYU Wagner Graduate School of Public Service, at the NYU Center for Urban Science and Progress, and a NYU Provostial Fellow for Innovation Analytics. She cofounded the Coleridge Initiative, whose goal is to use data to transform the way governments access and use data for the social good through training programs, research projects and a secure data facility. The approach is attracting national attention, including the Commission on Evidence Based Policy and the Federal Data Strategy.

Ekaterina Levitskaya is an Associate Research Scientist at the Coleridge Initiative, New York University. She utilizes computational approaches to the social science research, with special focus on text analysis and natural language processing. Her background is in computational linguistics and applied data science. She is interested in applying computational skills for the projects with social impact and utilizing text as data in a variety of applications for the social science research.

Ee-Peng Lim
Dr Ee-Peng Lim is the Lee Kong Chian Professor of Information Systems and Director of Living Analytics Research Center in the Singapore Management University. He received his PhD degree in Computer Science from the University of Minnesota. His research expertise covers social media mining, social/urban data analytics, and information retrieval. He has published more than 90 international journal papers and 280 conference papers, many of them appeared at top ACM and IEEE journals and conference venues. He is the recipient of the Distinguished Contribution Award at the 2019 Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD).

Jonathan Morgan is a Doctoral Candidate at the University of Mannheim. Jonathan has worked as a Senior Research Scientist at New York University; a Senior Data Scientist at the United States Census Bureau; a programmer, designer, and product manager for higher education systems integrations and data governance applications at various companies and institutions; and as an online producer for The New York Times and Multiplatform Editor for the Detroit News. He has a Bachelor of Arts in Computer Science from Wittenberg University, a Master of Arts in Journalism from NYU, and was a University Enrichment Fellow at Michigan State University.

Ian Mulvany is head of transformation at SAGE Publishing. He helped setup SAGE’s methods innovation incubator SAGE Ocean following a lean product development approach. Previously he ran technology operations for eLife, was head of product for Mendeley and ran a number of early web2.0 products for Nature Publishing Group. He is passionate about creating digital tools that support the research enterprise. He is interested in the interplay between different stakeholders that can lead to the sustainably of these kinds of tools.

Paco Nathan is a technologist, consultant, and evil mad scientist with deep experience in the areas of machine learning, human in the loop patterns for AI, and natural language work. He is advisor to several tech organizations including: NYU Coleridge Initiative, IBM Data Science Community, Amplify Partners, Recognai, Data Spartan, and Primer AI. He is co-chair for Rev conference by Domino Data Lab.

Axel-Cyrille Ngonga Ngomo is a full professor at the Computer Science department of Paderborn University. In his work, he focuses on the life cycle of knowledge graphs. He has hence been involved in the development of approaches for the extraction, storage, querying, integration and fusion as well as the exploitation of knowledge graphs. One core usage of knowledge graphs he explores is the development of explainable and responsible active machine learning algorithms. Axel is a proponent of open data, open research, and open science with a keen interest in paradigms and frameworks for reproducible scientific research.

Wolfgang Otto is a postgraduate and research associate at GESIS - Leibniz Institute for the Social Sciences in Germany. As part of the Knowledge Technologies for the Social Sciences Department under Stefan Dietze, he applies NLP-techniques on text and data corpora in the Social Sciences. After finishing with a master degree at the NLP Group at Leipzig University (Prof. Dr. Gerhard Heyer), he is part of a team in a third funded project (German Research Fund) to build up a

Specialized Information Service for Political Scientists (pollux-fid.de). A Project the State and University Library Bremen (SuUB) is realizing in cooperation with GESIS. During his studies, he collaborated in Projects on Digital Humanities, Applied Text Mining, and Data Science.

Sophie Rand is an Associate Research Scientist working on the Rich Context project at the Coleridge Initiative.
Previously, she was a Public Health Data Analyst at the New York City Department of Health and Mental Hygiene, first in the Bureau of Primary Care and Prevention, where she worked with data from Health Information Exchanges and Electronic Health Records in support of clinical-community public health programs; and in the Division of Disease Control, working with real-time Emergency Department, reportable infectious disease, and school health data. Sophie holds a Bachelors of Science in Engineering from the Cooper Union and a Master's in Public Health from the CUNY School of Public Health.

Michael Röder is a research associate and a Ph.D. candidate in the Computer Science department of Paderborn University. His research focuses on data gathering, data analysis and benchmarking of linked data systems. He has been involved in several research projects and reviewed papers for different scientific journals and conferences.

Haritz Puerto San Roman is a master's student in the School of Computing at KAIST and research assistant in IR&NLP Lab at KAIST. He graduated from the University of Malaga with a degree in Computer Science in July 2017. His research lays in the application of Machine Learning to Natural Language Processing, specifically to solve the problem of Question Answering.

Amila Silva
Amila Silva is a graduate from University of Moratuwa, Sri Lanka, with a First-Class Honors degree in Electronics and Telecommunication Engineering, where he was placed second of the graduating class of 110 students. He is currently working towards a Ph.D. degree at the Department of Computing and Information Systems, the University of Melbourne, Australia. He was awarded the Melbourne Graduate Research Scholarship supporting his studies. Besides, he was awarded the Rowden White Scholarship, a prestigious scholarship provided by the University of Melbourne to talented Ph.D. students. His research interests include continual learning, graph analytics, and data mining.

René Speck is a research associate and a Ph.D. candidate in the data processing service center (Research and Development Department II) at Leipzig University. His work and research focus on knowledge extraction, knowledge graphs, natural language processing, and machine learning. René Speck has been involved in several projects at the Leipzig University since 2013. He has been a reviewer for several conferences and journals since then as well.

Nikit Srivastava is a master's student and a student research assistant in the Computer Science department at Paderborn University. His research mainly focuses on data science chatbots and word embeddings. He has been involved in the development of many proofs-of-concept and prototype demonstrations for different scientific research papers and conferences.

Narges Tavakolpoursaleh is a postgraduate and research fellow at GESIS - Leibniz Institute for the Social Sciences in Germany. At the moment, as a part of a team, she involves in a third-party funded project (STELLA) that aims to create an evaluation infrastructure for search and recommendation services within productive web-based search systems with real users.

Ricardo Usbeck is a senior (guest) researcher at Paderborn University focusing on data extraction and information retrieval. His main interest is the combination of machine learning, statistics, and linked data. Ricardo is leading and executing several national and international research projects concerned with searching large amounts of heterogeneous and small, specific datasets using natural language.

Daniel Vollmers is a research associate and a Ph.D. candidate in the Computer Science department of Paderborn University. His research focuses on Question Answering knowledge extraction and machine learning. He has been involved in several research projects in these domains.

Alex D. Wade is Program Manager for Knowledge Graphs and Open Science at the Chan Zuckerberg Initiative. Alex earned his master's in library science from the University of Washington and has worked for the libraries at the University of California at Berkeley, the University of Michigan, and the University of Washington. Alex has spent his post-academic career working on problems in information retrieval, knowledge representation, and open science at Microsoft, Amazon, and Facebook, and currently works on the Meta service and the Open Science group at the Chan Zuckerberg Initiative.

Tong Zeng is a Ph.D. candidate in the School of Information Management at Nanjing University and a visiting scholar in the School of Information Studies at Syracuse University, working with Professor Daniel Acuna in the Science of Science and Computational Discovery Lab. Tong's research interests lie within text mining and scientometrics. In particular, he is interested in applying natural language processing and network science techniques on scientific literature to investigate, understand, and facilitate various aspects of scientific communication. His recent projects involve detecting dataset mentions in full text, assigning credit to datasets, and disambiguating authors at scale.

Andrea Zielinski is a Senior Research Scientist at the Fraunhofer Institute for Systems and Innovation Research (ISI), Karlsruhe, Germany and conducts applied research in Machine Learning and Text Mining at the Innovation System Data Excellence Center (ISDEC). She studied Computer Science with a focus on Artificial Intelligence and Linguistics at the University of Hamburg. In 2002, she received her PhD in Computational Linguistics from Saarland University. Since 2008, she also serves as a lecturer for Text Mining at the Department of Computational Linguistics, Heidelberg University, Germany. Her research interests lie at the intersection of Natural Language Processing and Machine Learning, particularly on areas relating to Text Mining and Semantics.

Chapter 1 - Introduction

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.
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).

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 in the Rich Context Competition was from Allen AI 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 (from Facebook) and the more recent Ray (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]{.underline}](https://coleridgeinitiative.org/richcontextcompetition#howtoparticipate))
- **October 15 2018:** Participants notified and Phase 1 data provided (see [[First Phase Participation]{.underline}](https://coleridgeinitiative.org/richcontextcompetition#phase1participation))
- **November 15 2018:** Preliminary algorithm submitted (see [[Program Requirements]{.underline}](https://coleridgeinitiative.org/richcontextcompetition#programreqs))
- **December 1 2018:** 15 finalists selected (see [[First Phase Evaluation]{.underline}](https://coleridgeinitiative.org/richcontextcompetition#phase1evaluation)) and Phase 2 data provided (see [[Second Phase Participation]{.underline}](https://coleridgeinitiative.org/richcontextcompetition#phase2participation))
- **January 15, 2019:** The algorithms of up to 6 teams are selected for final submission (see [[Second Phase Evaluation]{.underline}](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]{.underline}] (https://coleridgeinitiative.org/richcontextcompetition#phase2evaluation))

All the information provided to participants was available here

https://github.com/Coleridge-Initiative/rich-context-competition

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).

Chapter 2 - Bundesbank

# Conceptual issues when characterizing data in empirical research

Building blocks for user-centric data-services: Usage data to support empirical research

Stefan Bender[1], Hendrik Doll[1], Christian Hirsch[1] [2]

Empirical economic and social science research uses microdata for analyses to connect theory to socie-tal problems. We present conceptual lessons learned from a machine learning competition held to au-tomate the discovery of datasets, research methods and fields in these research publications. Obtained information from the competition can be used to inform the debate about added value of the used (micro) data. Being able to measure societal benefits of data access is important to put funding decisions on an objective basis, since much research data is generated by publically funded researchers or available from official institutions. The obtained information from the competition can also be used to build a user-centric dataset recommendation system. Both of these outcomes will elevate the current knowledge generating process of empirical research.

Introduction

Policy makers increasingly recognize that informed decision-making requires data on the characteristics of units of a population, such as individuals, households, or establishments (i.e. microdata). Only microdata can uncover interdependencies between entities and document disparate global develop-ments. Making microdata available for independent research is subject to legal requirements that are designed to prevent the disclosure of information concerning an individual person or business entity. At Deutsche Bundesbank, the Research Data and Service Centre (RDSC) is tasked with making microdata available for independent research while simultaneously ensuring statistical confidentiality.

To strengthen effective quantitative research through optimal microdata usage, the RDSC has engaged in a series of projects that are targeted at enhancing user experience. One specific project currently pur-sued by the RDSC is the development of a microdata recommendation system, which is based on how microdata is being used in empirical research. Describing microdata from the usage in publications dis-tinguished this approach from traditional metadata for researchers, which is largely based on how data is produced.

Empirical research papers are an obvious source of information on dataset usage. A useful microdata recommendation system needs to rely on a corpus of dataset usage, as large as possible. Hand-curating such a sufficiently large corpus is prohibitively labor-intensive and error-prone. Thus, the competition with results described in this book to automate the discovery of datasets, associated research methods and fields in social science research publications lays the groundwork for any future implantation of such a recommendation system.

In this chapter, we present lessons learned from the competition described in this book. We do this from a background of all authors in social science. Readers interested in the more technical aspects of the competition may find Chapters 6 to 11 helpful, where participating teams explain their approaches in more detail. Furthermore, while our lessons learned revolve around datasets, fields, and mentions, Chapter 5 discusses the operational approach as well as lessons learned for designing such a competition. Finally, Chapter 13 presents in more detail a potential framework for implemeting a microdata recommendation system.

Extracting dataset citations from publications is a fairly difficult task because of the variety of dataset ci-tation formats and the absence of training data. Besides empirical research support, the gained infor-mation is the basis to provide value for policy purposes in the G20 context. For example, by providing researchers with information about the use and availability of microdata previously not being available in a systematic way, the results of the competition described in this book and the ensuing microdata recommendation system are a step towards reducing data gaps that have been diagnosed in the aftermath of the financial crisis.

On a broader level, the outcome of the competition contributes to the ongoing digitalization efforts of the Deutsche Bundesbank. Extracting relevant information from research papers as an unstructured data source broadens the value of unstructured, underexplored, data. Thus, the results of the competition presents a well-defined use-case to turn tacit knowledge into codified knowledge by converting text into relatively well-structured information. As a concrete first institutional implementation of competition results, microdata-based research will be supported by turning unstructured information into a useful source of reference for researchers.

Insights from a research data centre perspective

Background on research data centres

Reseach data centres (RDC) present an established operational approach to facilitate access to confidential microdata for statistical purposes. This approach is based on the theoretical framework of the “Five Safes” which was initially developed by Felix Ritchie at the UK Office of National Statistics in 2003. [3] The first dimension refers to safe projects. This dimension mainly refers to the whether the intended use of the data conforms with the use specified in legislations or regulations. For example, a legislation may specifically allow users to use the data only for independent scientific research.

Safe people, the second dimension of the Five Saves framework, requires data users to be able to use the data in an appropriate way. A certain amount of technical skill or a minimum educational levels may be required to access the data. In contrast, safe data refers to the potential to de-identifying individuals or entities in the data. Safe settings relate to the practical controls on how the data is accessed. Different channels may exist which in turn may depend on the de-identifcation risk. In practice, the lower the de-identifcation risk the more restrictive the setting will be. Lastly, safe output refers to the risk of de-identifcation in publications from confidential microdata.

In response to the increased internal and external demand for microdata and the data confidentiality re-quirements, in 2013 the Bundesbank set up the Integrated Microdata - based Information and Analysis System (IMIDIAS) and established the Research Data and Service Centre (RDSC) (for a detailed moti-vation, refer to Kalckreuth, 2014 and Bender and Staab, 2015). At the RDSC of the Bundesbank, many of the technical and organizational measures put in place to protect confidential microdata follow the Five Safes framework.

A main principle of the RDSC is to give free access to Bundesbank micro data for independent research following the Five Safe approach. Motives for doing so are to stimulate cooperation projects between researchers inside and outside of Bundesbank, get feedback on relevant topics for Bundesbank (use published research results to increase the internal knowledge) and to strengthen evidence-based policy-making on for Bundesbank relevant topics. To fulfill these tasks, the RDSC has to ensure microdata is used effectively by providing excellent services. Implementing potential for structured feedback from researchers back to data production and new research enables an improving empirical knowledge generating process.

The data access provided by the RDSC of Bundesbank and the underlying legal requirements are described in detail by Schönberg (2018). In a nutshell, the requests of users to use microdata are first reviewed pursuant to legal requirements. After the application of a researcher is approved by the RDSC, researchers conduct their research project in a secure environment designed to ensure ongoing compliance with internal data policies and external government regulations. For most microdata this requires researchers to be physically present at the premises of the RDSC in order to analyse the data. Furthermore, only strictly anonymized research outcomes may be used outside of the secure environment.

The RDSC provides access to anonymized datasets on banks, securities, investment funds, enterprises and households, all of which can be accessed at dedicated researcher workstations or for most of the Bundesbank's surveys – as for the Panel on Household Finances (PHF) study – the RDSC offers so called scientific use files. In addition, the RDSC provides information and advice to researchers on data selection, data content and analytical approaches. Together with the relevant statistical experts, it ensures that the microdata provided are documented in detail and archived. In doing so, the RDSC works according to globally rec-ognized standards and was accredited as a research data centre (RDC) by the German Data Forum ("Rat für Sozial- und Wirtschaftsdaten").

The knowledge generating process of empirical research in the RDSC

In this section we present the knowledge generating process of empirical research in the RDSC of Bundesbank. Figure 1 depicts this knowledge generating process which can be organized along the four key dimensions (i) data services, (ii) research, (iii) publication, and (iv) (structured) user specific knowledge. For simplification, we assume, that data services effects research and that the outcome of research is a publication. From a publication the RDSC (or Bundesbank) is able to destill user specific knowledge, which can be used for better services (to the "next generation" of users). In the following we discuss these dimensions in more detail and show how leveraging on (structured) user specific knowledge (dimension iv) may elevate the knowledge generating process to a higher level.

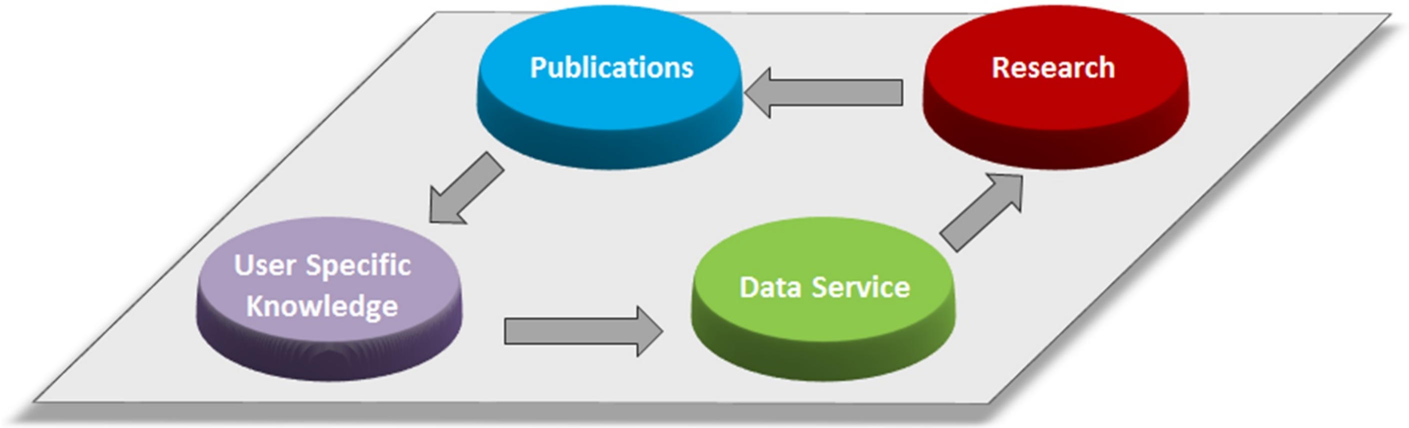


Figure 1: The four dimensions of the knowledge generating process of empirical research in the RDSC

The knowledge generating process starts from data sevices which are offered by the RDSC to researchers, who are analyzing data from the RDSC. Data services comprises raw microdata and comprehensive documentation of the data both of which the RDSC compiles together with the data-producing units in Bundesbank. [4] Furthermore, the data services dimension also includes the methodological improvement of microdata through e.g. applying record linkage techniques to facilitate the creation of new datasets for research. Finally, the RDSC also offers advisory services to potential and existing microdata users on topics such as e.g. dataset selection or analytical options.

The second pillar of the knowledge generating process in the RDSC is research. Researchers conduct their research project in the RDSC's secure environment and produce research outcomes. These outcomes – after statistical disclosure control by the RDSC – sometimes take the form of publications, which present results to the interested public in a form optimized for human consumption as unstructured text. These publications contain knowledge accumulated by researchers about data usage over time (experience), e.g., knowledge about dataset particularities, which in turn could be utilized to inform the debate on how to improve data services.

Examples of user specific knowledge acquired by researchers include:
+ How data is used (e.g. additional data cleaning, variable transformation, combining datasets, using additional information)
+ What purposes data is used for (e.g. topics, methodology, research area)
+ What kinds of analyses or techniques have been tried and are used ultimately
+ What information about data is most valuable to get to the results, respectively which linkage or data enrichment renders the data most valuable.

Being able to access structured user-specific knowledge through the competition described in this book enables improving data services by making discovery of data and related projects, people, and publications at Bundesbank more comprehensive and efficient. For example, knowledge harvested from publications may be used to enhance services provided by RDSC by allowing standard datasets to be tailored to the needs of re-searchers. Similarly, data producers benefit from feedback on their data, allowing them to improve data quality.

The challenge is to establish such a feedback loop. If effective feedback is given and used, the microdata-based knowledge-generating process restarts with data services, but it is elevated to a higher level. The effect of being able to leverage on information from this feedback loop is depicted in Figure 2. Better data services in turn allow better research, because available microdata is better customized and more effectively used. Better research will be published in higher ranked journals and will generate more relevant user specific knowledge, which can be used again for better services. To sum up, establishing a knowledge generating process as we have shown will lead to improvements in the four key dimensions of this process of empirical research.

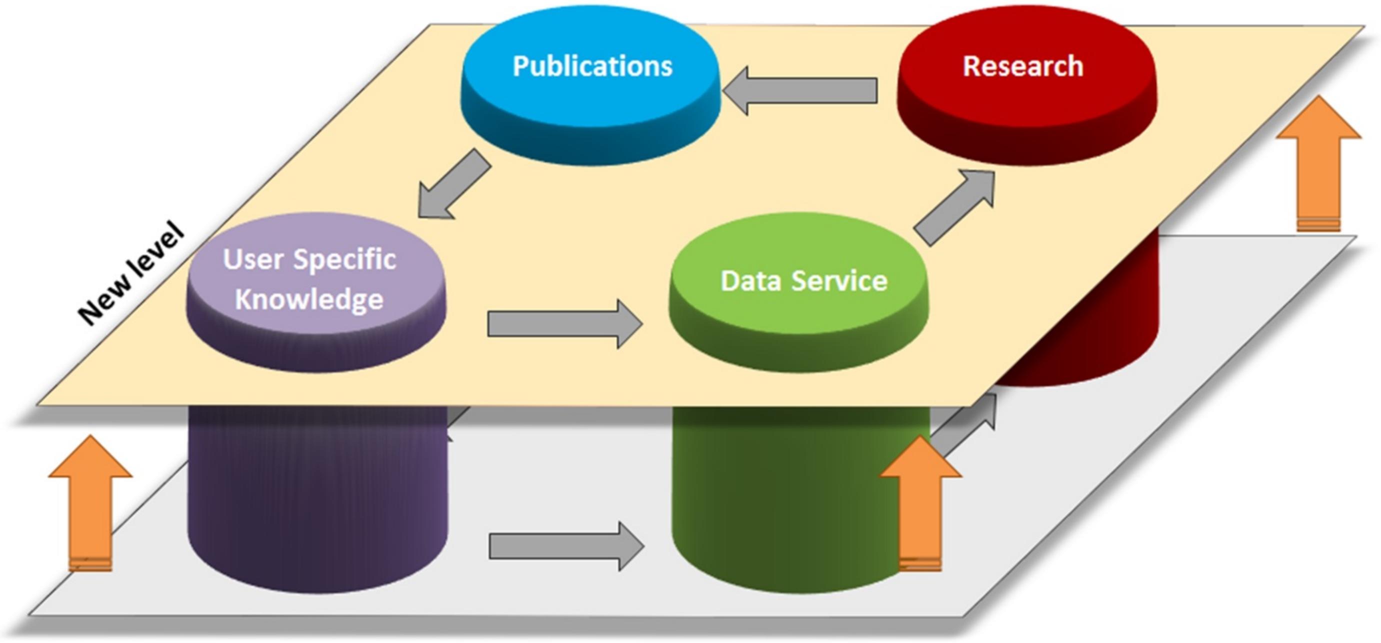


Figure 2: Elevating the knowledge generating process of empirical research in the RDSC to a higher level by enabling a feedback loop to data services.

At the moment, this feedback loop is not present in a systematic way. The aim of the competition presented in this book is to identify appropriate procedures to close the gap between publication and data services, which would enable transforming knowledge available in publications into generally re-usable knowledge to inform stakeholders (data producers, RDSC, decision makers at Bundesbank). The results of the competition will thus ultimately enable better data services which in turn will make research outcomes more efficient through the channel of a better data usage.

Added value of structured user-specific knowledge

This section details two applications of obtaining dataset usage information from publications that would add value to the data services provided for the RDSC. First, existing applications can be optimized in a user-centric way which would lead to obtaining refined products (e.g. improved researcher recommendations and data documentation). Second, the case for societal investment in free data access can be empirically fortified. Positive externalities (i.e. research as a public good) suggests a less then societally optimal provision of research data and related services. Obtaining a dataset impact factor can then make the case for further investment in microdata provision by concretely showing a dataset's impact.

The structured user-specific knowledge produced during the competition may be used to inform the design of a dataset proposition system for researchers. By obtaining information on dataset usage in publications, data is for the first time available to construct indices on data set joint usability (and dataset maps to visualize such indices). Such an index connects datasets through actual use by researchers that combined data sets in the past. This enables recommendations, such as, "Researchers, who used dataset A, also used dataset B".

Going further, the usability index can be expanded into a measure, how well new datasets fit each other. Without needing joint dataset usage in past publications goodness-of-fit measures may be predicted based on dataset usage in the same field, using the same methods or by additional metadata similarity. This can be a valuable accelerator to effectively distribute new datasets in the research community. While both indices can be implemented using only information from the competition, extensions may enhance value to users which are based on other information such as current metadata.

When thinking about user recommendations, the example is set by large online platforms. These online platforms can recommend from two dimensions of information (excluding interaction for simplicity). First, data is available on a large number of observed purchases per customer, which enables statements like "since you like products A and B, you might also like C". Second, data is present on large numbers of observed customers per product, which enables statements like "users like you also bought".

In our setting, with the knowledge generating process of empirical research in the RDSC, we consider researchers and datasets. The universe of data users and researchers is decently large (i.e. the first dimen-sion), but per user, we only observe a limited amount of "dataset consumption" (i.e. the second dimen-sion). Hence, we have a decent chance of recommending based on other users behavior. However, we have only limited means of predicting a single users future datasets needs based on his past personal "dataset shopping" behavior.

However, we suspect a simpler underlying behavioral model of “data shopping” compared to shopping through large online platforms, because publishing with one dataset is not a casual purchase. Instead, it implies real commitment relating to being content with the purchase (less cognitive dissonance). Thus, we suspect that, compared to online platforms, less data points per person are needed, in order to make sensible recommendations. Also, in order to gain more of the rare information per user, we can fall back on dataset citations, i.e. “indirect data usage”, as outlined in Chapter 3 of the book.

A challenge in building a data-driven recommendation system is to make sure that recommended da-tasets are indeed feasible to use, i.e. constitute meaningful recommendations. Thus, besides information about datasets, additional information such as fields and methods is needed to be ingested into the system. This additional information essentially constitutes additional links between datasets that helps better align datasets. This is especially true in the finance domain where linking microdata is a common feature in empirical research.

Second, the RDSC as part of a public institution has a responsibility towards its principals, i.e. society. Granting data access free of charge for researchers should be backed by empirically measurable bene-fits of such data provision. Benefits from data usage can justify societal investment in free data access. However, measuring societal benefits through data access is not obvious at first glance. One possible starting point of approximating societal benefits of data access can be to measure the creation of knowledge [5] created by specific datasets.

One can argue that added value of providing administrative microdata is the marginal benefit relative to the second-best comparable commercial database, if such a database exists. Also, one can argue that a dataset, which enables causal evidence, adds more value to societal knowledge, compared to previously available datasets, from which only correlations could be deduced if an important goal is to inform the policy debate. However, both of these methods require identifying which empirical result from a publication can be attributed to which dataset.

Lessons learned from competition

Related literature

Extracting dataset citations from publications is a fairly difficult task because of the variety of dataset citation formats and the absence of training data (for a recent overview of data retrieval see Koesten et al., 2019). Boland et al (2012) propose a weakly supervised approach, using a pattern induction method for the detection of study references in full texts. They use a corpus of 259 publications from the Social Science Open Access Repository (SSOAR). They use a bootstrapping approach, starting with a small corpus of manually created training instances. The resulting system InfoLink now informs SSOAR.

Boland and Mathiak (2015) describe dataset extraction as a twofold task, finding dataset citation string and following entity resolution (match the string to the correct entity/ DOI). Concerning entity resolution, they report the difficulty of broad survey dataset citations that ignore data variability (such as years, versions, questionnaire variants, etc.), motivating a dataset taxonomy. Named dataset citations are often underspecified allowing identification of the survey but not of the precise dataset (which of multiple sub-samples, aggregation levels, survey modes, etc.).

Zhang et al (2016) use a bootstrapping approach to extract dataset citations from 116 computer science journals publications. Ghavimi et al. (2016) use a similar approach for social science papers finding datasets with well-documented metadata. According to them, only 25% of all dataset citations are given in the references, highlighting the unstructured citation culture for datasets. We advance from these with an environment with less available dataset metadata and a corpus of publications from a variety of fields for our purposes. To tackle this, we continue with a larger hand-curated annotated corpus.

Metadata schemas for datasets are available, such as the DataCite metadata schema and the da|ra metadata schema, which complies with the DataCite schema (Helbig et al. 2014). They offer dataset taxonomies and standardized citation propositions, however their categories do not optimally support automatic search and extraction, if no unique dataset identification (such as a DOI) is used. [6] In the con-text of central banks that provide microdata, recent progress has been made in the context of INEXDA. A metadata standard (in line with DataCite) has been developed (Bender et al. 2018) and datasets pro-vided by the RDSC are all DOI registered.

Improving dataset citation is high on the scientific agenda in recent years. This notably includes promot-ing widespread usage of persistent and unique dataset identifiers. As available datasets spread across a large number of databases, identification of datasets is important for reproducibility and to credit data creation efforts to incentivize data creation and publication (Lagoze and Vilhuber 2017, McMurry et al. 2017, Mooney and Newton 2012). If unique and persistent dataset identification in publications were available, Ball and Duke (2011) raise the idea of dataset impact factors with such information.

Dataset mentions

This section presents lessons that we learned throughout the duration of the competition described in this book. These lessons originated from our motivation presented in the previous chapter and the role that we assumed. This role was twofold. First, we contributed to the corpus of publications and metadata given to competition teams. Our contribution consisted of a corpus of tagged publications of Bundesbank working papers, some of which use microdata provided by the RDSC. Second, we contributed to both evaluation phases of the competition. [7]

We organize this section around the three sets of information that where the main focus of the competition: datasets mentions, research fields, and (statistical) methods used. We begin by describing our a priori expectation of what a dataset is. We did not delve into definitions of a dataset but rather considered it sufficiently defined for our purposes (as empirical social scientists and for the competition).

Since our approach depends on getting to know the user-perspective, we thought it plausible to let usage in empirical papers define a dataset for the purpose of the competition. Having a background in working at a large provider of financial data, we had a vague idea that all datasets would look like those the RDSC provides access to, which consist mostly of collections of structured data in matrix or database form. These datasets typically are defined by a name and with a well-defined scope, thus allowing clear citation, usually including a unique dataset identifier (such as a Digital Object Identifier, DOI).

Lesson #1: datasets fall into two broad categories

Since the corpus of publication used for the competition spanned different domains (like healthcare, education, and others), we quickly realized that our dataset image had an econocentric bias. In social science, we learned, datasets can be categorized into two broad categories for the purposes of extraction. First, there are named datasets, i.e. well defined, usually large-scale and publicized datasets (e.g. Compustat).

Generally, named dataset mentions are short strings in the publications, have commonly used abbrevia-tions (e.g. MMSR), and often containing institution name or name of commercial data vendor. Some-times (rarely, but increasingly) these datasets can be identified by a unique digital object identifier (DOI). These datasets are usually well-defined in scope and time, with formal documentation available. While data is usually collected with a specific purpose in mind, such datasets are be used across multiple pa-pers and research domains.

The second dataset category is what we call created datasets. By created dataset we understand da-tasets usually collected or built by authors of a publication for the purpose of analyzing one specific re-search question. Often, created data comes in the form of small-scale surveys, (structured) interviews, or randomized controlled trials, RCTs. Such data normally does not have a trademark name, but instead one or multiple paragraph descriptions in the publication. Dataset information is blended together with information on data collection and sampling methods. Data reference at its most condensed form then comes in a structure like “we interview a given number of participants in a given region suffering from a given disease and code responses in the following way”.

In contrast to named datasets, created datasets usually are not referred to by a specific string or com-monly used abbreviation. Data collection is usually paper specific, and the universe of existing datasets are not easily searchable. This makes it hard for text mining algorithms to correctly extract strings referring to dataset entities. Specific created datasets are harder to use for follow-up research, and reproducibility is given only if publishers provide data together with the paper. Therefore, the lack of unique identification and search terms renders data collection potentially redundant and dataset spread not optimal.

Lesson #2: Fractions of dataset category are domain specific

Throughout the competition it became clear that the fraction of named and created datasets varies across social science domains. Since different fields of social sciences rely on different identification techniques and differing potentials for conducting RCTs, the predominantly used data sources naturally vary. This has important repercussions for designing a competition, since algorithm performance and later recommendation system performance varies with the input corpus and the application field.

The number of datasets used per empirical paper (linked data) also varies across research areas. This number is also dependent on named vs. created datasets. In fields with widespread use of multiple datasets at once, the added value of recommending additional useful data might be expected to be higher than in fields that create study-specific data every time. Conversely, one could argue that the marginal utility of adding additional datasets is decreasing.

The optimal way forward is to start a data recommendation system for research field with higher expected marginal utility from additional datasets. In our view, these are research areas with widespread usage of named datasets. Named datasets are constructed without the concrete research question in mind. That is why information to answer a particular research question often has to be obtained from more than one data source and is particularly true in empirical economic and finance research.

Lesson #3: Unique identification of datasets remains an issue

From the distinction above, one could make the argument that named datasets are easier to identify than created datasets. However, this is not the case, because the same dataset name can refer to multiple subsamples or waves of same datasets, and it is unclear where to make distinctions between dataset entities. This makes it difficult to identify the mentions referring to the same data points. Issues are, just to name a few, different time periods or subsamples, different states of data and states of knowledge, computational data pre-processing or enrichment steps. These identification issues render the current task of entity resolution of extracted dataset mentions complicated.

Unique dataset identification carries significant repercussions for reproducibility purposes, where identi-fying the exact data used for a study is paramount. For reproducibility purposes, the current solution to this dataset identification problem is the direct data upload to the publisher together with the publication. This is neither storage-efficient for large datasets nor feasible in the case of confidential microdata. A more flexible way to solve this issue is to assign unique identifiers (DOIs) to the datasets.

With a DOI (identifying the exact time frame, sampling universe, data version, wave, aggregations, state of knowledge, etc.), datasets are identified and quantitative research using confidential microdata is re-producible. To make lives easier, DOIs also drastically facilitate the automatized extraction of well-defined datasets from publications (comparable to largely standardized citations of other publications, allowing easy retrieval of publication networks, etc.).[8]

[9] An alternative approach to ensure identification of datasets could be providing richer, more systematic metadata for datasets. See Chapter 4 in this book for a detailed discussion of this point.

Summarizing, if we successfully identify datasets and solve the issue of entity resolution, we can link and propose created datasets and thereby enable further research with such data, which takes up a notable fraction of publications in certain fields. While this task is harder than for named datasets, the potential for improvement remains larger as of today. For created datasets, too, DOI usage would be desirable; however encouragement or enforcement to use DOIs is harder in this case, because of a larger target group – authors instead of a limited number of data stewards. Even in case of widespread DOI usage for named datasets, the competition algorithms yield valuable results through the created datasets extraction in order to allow referencing and making available datasets used in the past for further analysis.

Lesson #4: Datasets mentions could indicate used for analysis vs. cited

After a discussion about dataset types and usage in fields, the last lesson that we learned about datasets concerns the mention of datasets in publications. These mentions come in two types. First, datasets used for empirical analysis and second, cited datasets in the literature review or references. Dataset citations (without empirical usage) can generally occur in the literature review section, even in theoretical, methodological papers, e.g. a given paper might report summary statistics based on datasets (“Author Y uses Compustat to…”). Sometimes differences between cited and used datasets are only semantic in nature. In well-written papers, the difference is usually fairly easy to distinguish for humans, but less clear for algorithms.

A key lesson we learned, is to think ahead of time, what the informational need is for the use-case at hand, used or cited datasets. Note that in an optimal setting, if information were available on the universe of datasets used for analysis in papers and on all publication citations, dataset citations would be redundant. This comes from the fact that a dataset citation in one publication is based on a dataset used for analysis in another publication and can be linked via available literature citations.

While literature citations are mostly standardized within research domains and are relatively straightforward to extract (hence publication networks / publications maps exist), information on used datasets in papers remains incomplete (even after the competition). Because of this, for the competition, we asked for used and cited datasets. It is important to note, that extracted dataset citations are always incomplete, since some authors report aggregate statistics from a different paper, but not the data behind (“Smith et al show…”).

If well separated, through extracted dataset citations, one obtains a “dataset map”, thus the “closeness of datasets”, and network measures such as centrality distinguishing important datasets (“nodes”). Through extracted empirical dataset usage on the other hand, one obtains relevant information for our purposes, namely information relating to dataset similarity and joint usage possibilities from the user perspective. However, for our envisioned recommendation system, usage of cited data (“indirect” data usage) is a valuable feature, since it yields more limited data on dataset “purchases” of a user.

As training data for the algorithms it is important to include theoretical literature, essays, etc. in the corpus of publications. [10] Obviously, this is helpful for algorithms to correctly identify true negatives, i.e. correctly identifying theoretical papers. For this task, distinguishing between cited and used datasets becomes relevant once again, because clearly separating theoretical papers that merely cite data from empirical papers depend on such a distinction.

[11] Chapter 12 discusses potential strategies for collecting training data in more detail.

Fields and Methods

The competition also asked participants to extract information about research fields and methods used in the publication. We want to gather this information from the user side, because data producers and annotators do not necessarily foresee all usage potential for their data and the point of our envisioned system is to increase user value. One such idea is to construct dataset similarity indices from the usage side, information is relevant not only on existing joint usage by others ("people like you often used dataset Y, too" – hence dataset extraction), but also on new dataset or linkage potentials ("this might also interest you based on your preferences"). For this, information is necessary on the context, how datasets are used.

Lesson #5: Think before you act: define fields and methods

To obtain the most relevant categories of research fields, we did not provide any thesauri to the competition, on purpose. The rationale behind this was to see the unhindered creativity of teams, which available information sources they would use or not use (e.g. reference datasets, Wikipedia, archive.org, other repositories, thesauri, statistical clustering techniques, etc.). On the other hand, thesauri limit the catalogue of potentially identifiable fields and methods, thus prohibiting new methods and fields to be identified in fast-changing modern research areas. Also thesauri might disturb algorithm performance, since algorithm might be forced to categorize topics and fields to older or less exact categories than necessary.

However, using thesauri does have well-known advantages, as any librarian will confirm. These advantages include easy clustering of similar fields and methods and a manageable category set of predictions. For field predictions, we generally face a fine line between too broad predictions (safe, but unin-formative) and too narrow predictions (narrow, but potentially wrong). A potential way out is backward induction here – we can present differently aggregated predictions for fields to users and get feedback from them (let users rank usability – "Was this helpful to you?").

Concerning our definition of methods for the purpose of the competition, two questions arise. The first is the definition of statistical methods (i.e. inclusion of sampling methods, qualitative methods, etc.). Secondly, there are multiple statistical methods in a publication (besides the main causal analysis, there can be methods reported for data preparation, sampling, baseline results, robustness checks, descriptive statistics, etc.) and issues of potential weighting of importance of these.

For useful new recommendations to be provided to researchers, we decide to include in statistical methods all methods that describe potential for a merge of datasets / joint usability, hence to include all the above listed. We consider a broad definition of methods, not only including high-level statistical methods, such as ordinary least squares, but also including the observed unit, time period or even re-gression equations. If two papers then use different datasets in the same field using the same methods, there is a relatively high likelihood that those datasets can be linked or used together to create new in-sights.

Discussion

Several decades ago, publication citation networks were constructed and to our knowledge no such undertaking has yet been done for datasets. This comes from the fact that no curated training data corpus is readily available in decent quality. Since no such data is available, we manually annotate papers for the competition and now propose to go forward with this in a larger scale.

We would have no need for this competition in a world with universal dataset identifier usage (such as DOIs). In such a scenario, unique identification and standardized citations of datasets would be readily available. Since DOIs only now and slowly gain widespread application for datasets in social science, our task is a 1:n mapping of publications to datasets without unique identifiers. For scientific papers, many journals already provide DOIs for papers.

There are ongoing efforts by journals to have all used data published for reproducibility reasons. Incentivizing researchers to provide unique identification of datasets used in papers is a logical next step. This will ensure reproducibility for confidential microdata and facilitate our use-cases. In the meantime, we show a way forward to learn from the current state of information and analytically use presently available information.

The competition highlights that datasets can be categorized in different dimensions for the purposes of extracting dataset mentions from publications. We propose a binary distinction of datasets into named as opposed to created datasets. As named datasets, we consider formal, large datasets by commercial or official institutions, often referenced in relatively standardized forms as commonly used abbreviations. Created datasets are those created for the specific purpose of one research question in mind. They are generally described in less standardized paragraphs. Usage of named versus created datasets varies across research areas.

Also varying across research areas is the number of datasets used per empirical paper. This number al-so depends on the spread of formal, named datasets as opposed to created datasets for single studies. In fields with widespread use of multiple datasets at once (linked data), the added value of recommending additional useful data might be expected to be higher than in fields that create study-specific data every time. Conversely, one could argue that the marginal utility of adding additional datasets is decreasing. The optimal way forward is to start a data recommendation system for research field with higher expected marginal utility from additional datasets.

Conclusions

In the competition described in this book, we asked teams to extract datasets, fields and methods from a corpus of hand-annotated research publications. The value of the extracted information lies in informing a user-centric dataset recommendation system and thereby enabling optimal and timely spread of available datasets throughout the research community. Furthermore, such information allows us to compute dataset impact factors by obtaining data-driven information on which datasets underlie high-quality research outputs. This in turn is a proxy for societal benefits of data provision by research data centres, thus motivating in-vestment in data access infrastructure.

We introduce a circular model of the knowledge generating process, which increases in levels. From data services, research is conducted, publications are published and user-specific knowledge is generated. Having such knowledge on dataset usage, data services in turn can be improved, which elevates research, publications and user specific knowledge. Thereby, the circle repeats on a higher level. The current competition works on strengthening the knowledge pillar as well as the transmission mechanisms from publications to knowledge to improved data services. [12]

Automatic processing of generated knowledge in publications becomes increasingly available with modern text analysis tools. Extracting such information is important, because timely and optimal usage of gained results increases the speed, by which findings can be incorporated into data services and thereby next-level research is enabled in turn. To further improve automatic processing, minimum standards for dataset taxonomy are needed. Harmonized metadata schemas for data sets – like the INEXDA metadata schema for central banks and statistical offices (compliant with and building upon DataCite) – offer such an approach.

The competition showcased that information extraction of the necessary information for such systems is possible. The delivered prototype algorithms prove this claim. With the proof of concept, there is a more substantiated case for investing in a larger hand-curated training corpus of annotated research papers. On the road towards a user-centric dataset recommendation and metadata system, the competition forced us to clarify organizational needs and methodological aspects.

For the way forward, it is important to note the importance of the research area on the strategic path towards a unified user-centric microdata recommendation system. The choice of the research domain will greatly influence algorithm performance. Since human effort in creating training data is expensive, one should deliberately pick research domains to start with. This arises because text extraction algorithms (and humans) struggle with informally described created datasets. The low-hanging fruits of prototyping dataset recommendation systems, usability indices etc. are easier to implement for research areas with a largely formalized dataset citation culture (however ultimately potential for benefits may well be larger in other research areas).

References

- Ball, A., and M. Duke (2011): How to cite datasets and link to publications. Digital Curation Centre.
- Bender, S., Hausstein, B., & C. Hirsch (2018). An Introduction to INEXDA's Metadata Schema. Technical Report 2018–02, Deutsche Bundesbank, Research Data and Service Centre.
- Bender, S. and P. Staab (2015). The Bundesbank's Research Data and Service Center (RDSC), Gateway to treasures of microdata on the German financial system. IFC Bulletin 41 (2015).
- Boland, K., Ritze D., Eckert, K., & B. Mathiak (2012): Identifying references to datasets in publica-tions. Theory and Practice of Digital Libraries, pp. 150–161. Springer Berlin Heidelberg, http://doi.org/10.1007/978–3–642–33290–6_17
- Desai, T., Ritchie, F., & R. Welpton (2016). Five Safes: designing data access for research, Working Papers 20161601, Department of Accounting, Economics and Finance, Bristol Business School, University of the West of England, Bristol
- Ghavimi, B., Mayr, P., Vahdati, S., & C. Lange (2016). Identifying and improving dataset references in social sciences full texts. arXiv preprint arXiv:1603.01774.
- Helbig K., Hausstein B., Koch U., Meichsner J., & A. Kempf (2014): da|ra Metadata Schema. Gesis Technical Reports 2014/17, DOI:10.4232/10.mdsdoc.3.1
- Von Kalckreuth, U. (2014). A Research Data and Service Centre (RDSC) at the Deutsche Bundes-bank–a draft concept. IFC-Bulletin No 37, Irving-Fisher Comittee on Central Bank Statistics.
- Koesten, L., Mayr, P., Groth, P., Simperl, E., & M. de Rijke (2019): Report on the DATA: SEARCH'18 workshop-Searching Data on the Web. ACM SIGIR Forum (Vol. 52, No. 1, pp. 117–124). ACM.
- Boland, K. & B. Mathiak (2015). Challenges in Matching Dataset Citation Strings to Datasets in Social Science. D-Lib Magazine 21, 1/2.
- McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., & A. Gonzalez-Beltran, A. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS biology, 15(6), e2001414.
- Mooney, H, & M. P. Newton (2012): The anatomy of a data citation: Discovery, reuse, and credit. eP1035-eP1035.
- Schönberg, T. (2018): Data Access to Micro Data of the Deutsche Bundesbank. Bundesbank Tech-nical Report 2018–01.
- Vilhuber, L. & C. Lagoze (2017): Making Confidential Data Part of Reproducible Research. Chance
- Zhang, Q., Cheng, Q., Huang, Y., & W. Lu (2016). A bootstrapping-based method to automatically identify data-usage statements in publications. Journal of Data and Information Science, 1(1), 69–85.

Chapter 3 - Digital Science Use Cases

Chapter 3 – Digital Science Use Cases: Enriching context and enhancing engagement around datasets

Christian Herzog[1,a], Daniel W Hook[1,2,3,b], Mark Hahnel[1,c], Stacy Konkiel[1,d], and Duane E. Williams[1,e]

^1[Digital] Science, London, N1 9XW, UK

^2[Department] of Physics, Washington University in St Louis, St Louis, Missouri, USA

^3[Centre] for Complexity Science, Imperial College London, London, SW7 2AZ, UK

(^a^[[https://orcid.org/0000–0002–9983–0033]{.underline}](https://orcid.org/0000–0002–9983–0033),
^b^[[https://orcid.org/0000–0001–9746–1193]{.underline}](https://orcid.org/0000–0001–9746–1193),
^c^[[https://orcid.org/0000–0003–4741–0309]{.underline}](https://orcid.org/0000–0003–4741–0309),
^d^[https://orcid.org/0000-0002-0546-8257]{.underline}
^e^[[https://orcid.org/0000–0002–2111–3413]{.underline}](https://orcid.org/0000–0002–2111–3413))

3.1 Introduction

The relationship between research, researchers and data is changing. Data has always played a critical role in scientific research, however in recent years it has taken centre stage not only for the so-called hard sciences, but also for the social sciences, and it has an increasing role in the humanities (Giuliano, 2019). We assert that this change is, at least in part, driven by two key factors: First, the increasing volume of data that is available to researchers either, for example, through the increasing sensitivity of instruments that aid experimental work, or through the ubiquity of computer systems with which we interact in our daily lives. Second, our ability to process and analyse these data is growing quickly as computers become faster and algorithms become more powerful. While some researchers welcome having more data to work with, others are challenged or marginalised in this new data-rich world (Eijnatten et al., 2013; Grusin, 2014; Leurs and Shepherd, 2017). These effects are often compounded by the tools that researchers must master to connect their research to data.

In the hard sciences, the CERN Open Data Portal contains 131 datasets describing particle collisions, each of which comprise around 300Gb of data at the time of writing (CERN, 2019). In the social sciences, the CISER Data Archive at Cornell is home to more than 1000 different social science data sets (CISER, 2019). These examples are individual instances chosen at random from many that could be used to demonstrate the variety and scale of data available for research. However, even these examples don’t begin to quantify the amount of detailed personal data available to companies such as Facebook. Data of this latter kind has already been used in academic studies as well as in more controversial contexts (Jordan and Weller, 2018; Kamp et al., 2019; Stark, 2018). Clearly, there is an increasing diversity and depth of data available for research from both traditional and from new sources.

Many researchers now work with large volumes of data. Fortunately, many facilitating technologies have become commoditised and are available at a fraction of their original cost: storage is cheap and data transfer is fast. But, Increasing the value of data to researchers is no longer about technology, rather it is about the information and culture around the data.

In this chapter, we take our lead from Chapter 1 in recognising not only that science is at a crossroads but that the whole of research is changing. We discuss two elements of infrastructure that, if enhanced, can make data more useful and valuable to the whole research community: information infrastructure and cultural infrastructure. The Rich Context project supports the development of tools that enrich not only information infrastructure around datasets, but which also enhance the cultural infrastructure. *Information infrastructure* includes details of the approach to data stewardship, context of usage, code applied to the dataset in its production, as well as code applied to the data to derive further results or translate it for practical uses. All these factors add critical elements to the research infrastructure. *Cultural infrastructure* includes creating the incentives, triggers and frameworks that encourage the dataset stewards, experts and users to contribute to these critical information elements.

3.2 Information Infrastructure

Information infrastructure can be defined as the collection of processes and artefacts that are foundational to today's scholarly communications. A simplified model of scholarly communications would have artefacts such as journals, journal articles, article metadata and citations. In this case, the processes would be peer review and scholarly search.

When creating one of the first scientific journals *Philosophical Transactions* 350 years ago, the members of the Royal Society did not have today's data-centric world in mind. While a clear line can be drawn from the articles of that time to the articles of today, infrastructures have grown up around research publications in the intervening years that have moved the structures and expectations of the research article forward. These norms are powerful and persistent through their ubiquity. For example, in the large majority of modern research literature, we continue expect articles to be grouped into journals and published on a specific date, and we expect there to be a version of record that constitutes a definitive record of a piece of research.

Data is more fluid than a standard research article: it is produced and updated more frequently and iteratively; it needs to be shared with many in a collaborative context; it is processed and versioned by different colleagues. Data does not fit into the normalised research publication. Research fields that rely on data are beginning to publish data as a distinct output from a research article. Data is becoming a principal research output, while the technological challenges of publishing data are being addressed, the format and necessary fields of the metadata that describe data, the file format in which the data resides, the resource to annotate the data to make it useful to others, the way in which data should be cited in a paper or by another dataset, the description of the processing that has been applied to the data, the details of the ethical review process behind the exercise that gathered the data, and many other norms do not yet exist homogeneously across subjects and geographies. There are not yet strongly established norms that help researchers to have trust in data.

A dataset can change with time for many reasons: data may be added over time, corrections may be issued, and so on. In these cases, it may be appropriate to “version” the dataset (by issuing a persistent identifier for a point-in-time snapshot for the dataset, allowing subsequent changes to receive their own “versions”). But changes to a dataset may have a knock-on effect on the interpretation of the data and may fundamentally alter the research result that was originally reported. Moreover, in many fields “Big Data” is so central that it not only puts pressure on the community to establish an acceptable model of data publication, but also puts significant stress on how we read, interpret, and review research as a whole.

Many datasets are now so vast that we lack the ability as humans to consume them in an easy way. Visualisation technologies and other tools that allow us to interact with and sample data dynamically have received significant attention in recent years, and have helped with the interpretation of data in online environments. But it is simply impossible to reduce some types of data to a single figure or printable table, as would be the case for “traditional” journal publishing. By attempting to do so, we miss the essence of the data and risk failing to communicate data-driven conclusions accurately. This limitation of current publication formats (e.g. static PDF files for articles) is an issue that relates to the reproducibility crisis of modern research.

Peer review is another process that is not easy to apply to data as a “first class” research object. Historically, peer reviewers have ensured that a piece of research is well-communicated and correct in the sense that it is reproducible. This level of peer review is difficult to apply in the context of research data. If data is being published as a primary output, then it may be possible to perform a kind of peer review by applying some statistical tests to a sample of the data, or by using some other appropriate technique. However, it is no longer practical in most cases to set up a parallel experiment to reproduce data, as had been the case in years past. Across all contexts there are good reasons for these challenges: the experiment may be too costly to repeat, or the conditions of the original data collection may not be replicable (for example, surveying stress levels of the populace during a specific political event). In addition, ethical considerations such as the anonymity of those being surveyed may make certain types of data difficult to review. Thus, we need to develop robust and accepted approaches to peer review, not only for data itself but also for those publications that are heavily based on data. Without peer review or some suitable proxy for peer review that makes sense for data, it is difficult to know whether a dataset can be trusted. Without trust, a dataset has no value to a researcher who seeks to build upon it.

Several publishing innovations have made journal articles more discoverable and accessible in recent years, such as preprint servers, the widespread use of Digital Object Identifiers (DOIs), and centralized search engines. However, while some of these infrastructures do enhance a researcher's ability to find research data, they do not fully translate from the realm of journals to data. There are multiple reasons for this lack of translation, some of the key features include: a) weakness of a homogeneous metadata infrastructure for datasets; b) inhomogeneity in the types of data that can be shared; c) proliferation of different platforms to store data; d) lack of standardised publication practices; e) lack of adoption of standards across fields. When compared with the "shape" of an academic article for which there is a standard structure (e.g. DOI, abstract, title, authors, keywords, etc) specifically designed to facilitate human search, it is clear that datasets are contextualised by an immature information infrastructure.

Datasets are more complex to classify and annotate than articles, yet some progress is being made. The core fields required to create a valid DataCite record are identifier, creator, title, publisher, publication year and resource type (DataCite Metadata Working Group, 2016). All other data fields are optional (e.g. location, funder, subject, contributors) due to the fundamental uncertainty in what might constitute research data in the future. This flexibility limits how data can be discovered. It has taken some years for Web of Science, Google and others to introduce functionality to search for datasets in their discovery systems.

Technological infrastructure for data--or lack thereof--has huge implications for the discovery, peer review, citation practices, interpretation, and availability of data. These challenges are interconnected with challenges we face when thinking about the cultural infrastructure for data, as well.

3.3 Cultural Infrastructure

There are two main aspects to cultural infrastructure: incentives and capability. Both aspects affect how researchers engage with research data, and their behaviours relating to sharing it with others and making it available to external scrutiny.

Anecdotally, academics do not typically take up research careers for financial gain. Rather, they choose to dedicate themselves to understanding a specific problem or field partially in the hope of making a discovery. For most researchers, success is not strongly coupled to prize winning, but rather by winning the freedom to determine their own research agenda. Researchers in many fields are promoted by publishing in specific high-impact journals, leading to funding success, which in turn usually leads to greater control of your research.

Sharing data is often not well-aligned with the current model of incentives. Parting with the data that underpins your research gives rise to two concerns. Firstly, that someone may find an error in your work and discredit what you have done. Secondly, that someone else may not share their own data but will gladly reuse yours if you make it available. This is especially the case in fields where success is based on having more data to analyse.

A further level of inequity exists in which data-related jobs are valued by the Academy. If a researcher happens to be particularly talented in working with data curation, data analysis or data processing, there is no track for recognising these talents. They are unlikely to be a first author on a publication in a major journal due to their data wrangling talents, and hence they have less of a chance of career progression than researchers who take a more traditional “publish or perish” path with their work as described above.

This set of perverse incentives means that people with the capability to handle data are often incentivised to leave research. Hence, not only do we have a problem of incentives in sharing and communicating data, but we also have a problem in retaining people who have the capability that we need to structure data so that it can be shared and built upon.

Capability for sharing data is the second aspect of the cultural challenge that academia continues to wrestle with. Making data available to others is generally accepted as a key part of the research communication process. However, there are certain established norms around when the data should be shared, and to what depth it is shared (Linek et al., 2017); for example, in fields where human subjects research is prevalent, there is a much more conservative attitude towards open data than in fields like astronomy where data sharing is widely practiced, given that data can be collected by only a handful of observatories and telescopes worldwide.

In fields that are more applied, ensuring that data generated as a result of a commercial relationship is protected is crucial. In such fields, academics often have a better understanding of copyright, intellectual property rights and licences (Treadway et al., 2016). But outside of this context, there is a general lack of understanding of these issues and hence data are often not shared over concerns for a perceived legal barrier.

Other concerns are ethical—for example, should these data be shared if it might infringe the rights of the subjects of the research? Researchers are beginning to become aware that, through the use of algorithms, some data is not as well anonymised as it may first appear (Siddle, 2014). Anonymisation of data is a research field in and of itself (Li et al., 2007).

The degree and nature of ethical issues and industry-proximity vary greatly between different research fields and give rise to different cultures of data usage and re-usage across fields and even within fields. Some researchers are motivated to engage with the open access community and hence choose approaches to sharing data that include granting permissive licences, association of unique identifiers with data, adherence to data standards and training students to adhere to similar approaches. Other researchers are motivated to ensure that data are not shared due to the information that can be inferred by processing the data.

The power of the newest algorithms, or of algorithms yet to come, mixed with constantly developing ethical nuance means that it is difficult to pre-empt what may or may not be acceptable to share in the future. Hence, some may feel that it is simply better not to share, especially in the social sciences, where many of these issues are more prone.

Other concerns are simply practical—how does one make data available in a way that is meaningful to others? The work associated with making a dataset generically machine-readable is challenging for many researchers, who are not to be experts in data handling. The work associated with making a dataset human-understandable, reproducible and fully contextualised is often significant. Funding constraints may make it impractical to share data or to add useful, valuable or even critical annotations to a dataset. However, funders are beginning to prioritise these activities in their grant programs (Jisc, 2019; NNLM, 2019). All these factors lead to uncertainty exacerbated by different levels of confidence and understanding and consequently an uneven landscape in what is shared, how it is shared and where it is shared.

3.4 Enriching context

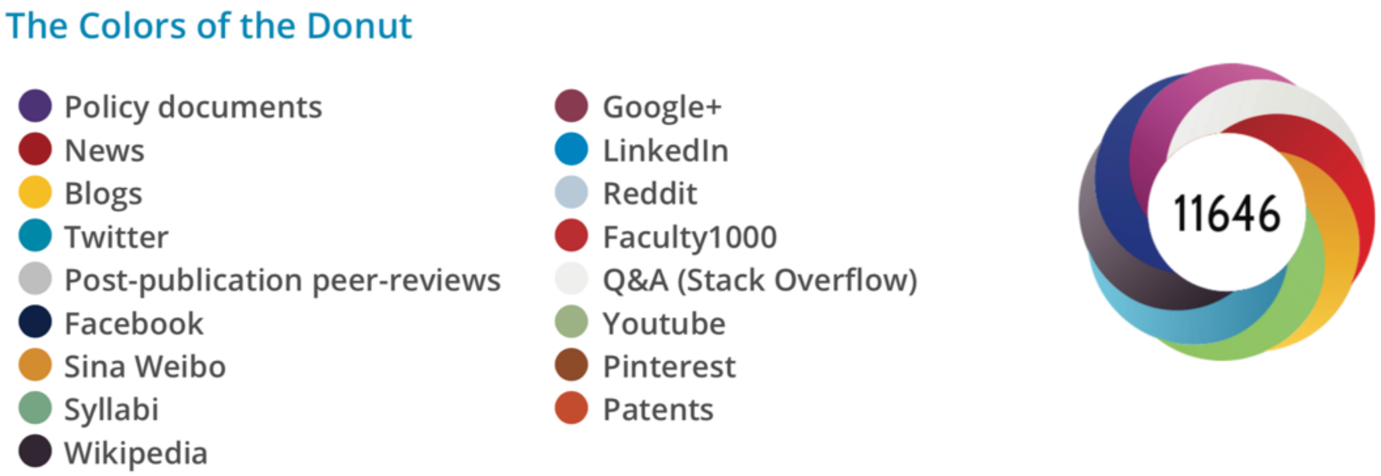
The points discussed above offer some indication of what would be needed to improve the value of research data. Firstly, to address issues of cultural infrastructure, we need to adopt an expanded version CredIT (Allen et al., 2014) that focuses on datasets. This expansion would ensure that all contributors to a dataset's creation, development and maintenance over time are stored in a machine-readable format. Such a record is central to the facilitation of culture change across research. Only with this structure in place can the activities around datasets be readily recognised and incentives created that would support data sharing. Secondly, to address the deficits in information infrastructure, a set of tools that allow research data to be discovered and contextualised needs to be introduced. In this section, we focus on this second challenge.

The ability to add context any piece of research was a strong driver for the creation of Dimensions (Hook et al., 2018). The idea that all research happens in a particular place, at a particular time, carried out by a set of people, some of whom may be affiliated with a research institution, gives a set of metadata that allows us the "weak context" of a piece of research. By "weak context" we mean that the context being provided gives no deep understanding of the context of an article to a non-expert and is essentially indistinguishable from standard metadata. But with modern data mining approaches, it is possible to add a "strong context".

Strong contextualisation of research should provide a user with rich information about the research including funding, other research produced as part of the larger project (e.g. related publications, clinical trials, etc), and details of the research that was built on top of it. This information should also fit into, trends and graphical representations that offer a more complete, more rapid understanding of how research fits into the larger field, related fields, or the context of the publishing journal or supporting institution. For example, for a research article, we should be able to quickly understand how many researchers are in a related field, whether the field is growing, how old the field is, how much funding has been deployed in the field, which countries have provided that funding, whether the field has begun the translation to application through patents or clinical trials, or whether it has been used as a basis for the formulation of policy.

Context can also be offered in the data that we provide to understand the reach and influence of research.

Alternative metrics ("altmetrics") are data from the social web that run orthogonal to classic citation measures, which can be seen to add significant context to an article – extending our understanding of how different cohorts of potential users of the research are engaging with it. For example, altmetrics can be used to understand if an article is being mentioned in the news, in which geographical regions it is being noticed, whether it is being used as part of a teaching syllabus, and many other kinds of public and non-traditional scholarly engagement. These data can then be visualized in creative ways to add instant additional context to engagement with a research article (see Fig. 3.1).



{width="7.270833333333333in" height="2.5277777777777777in"}

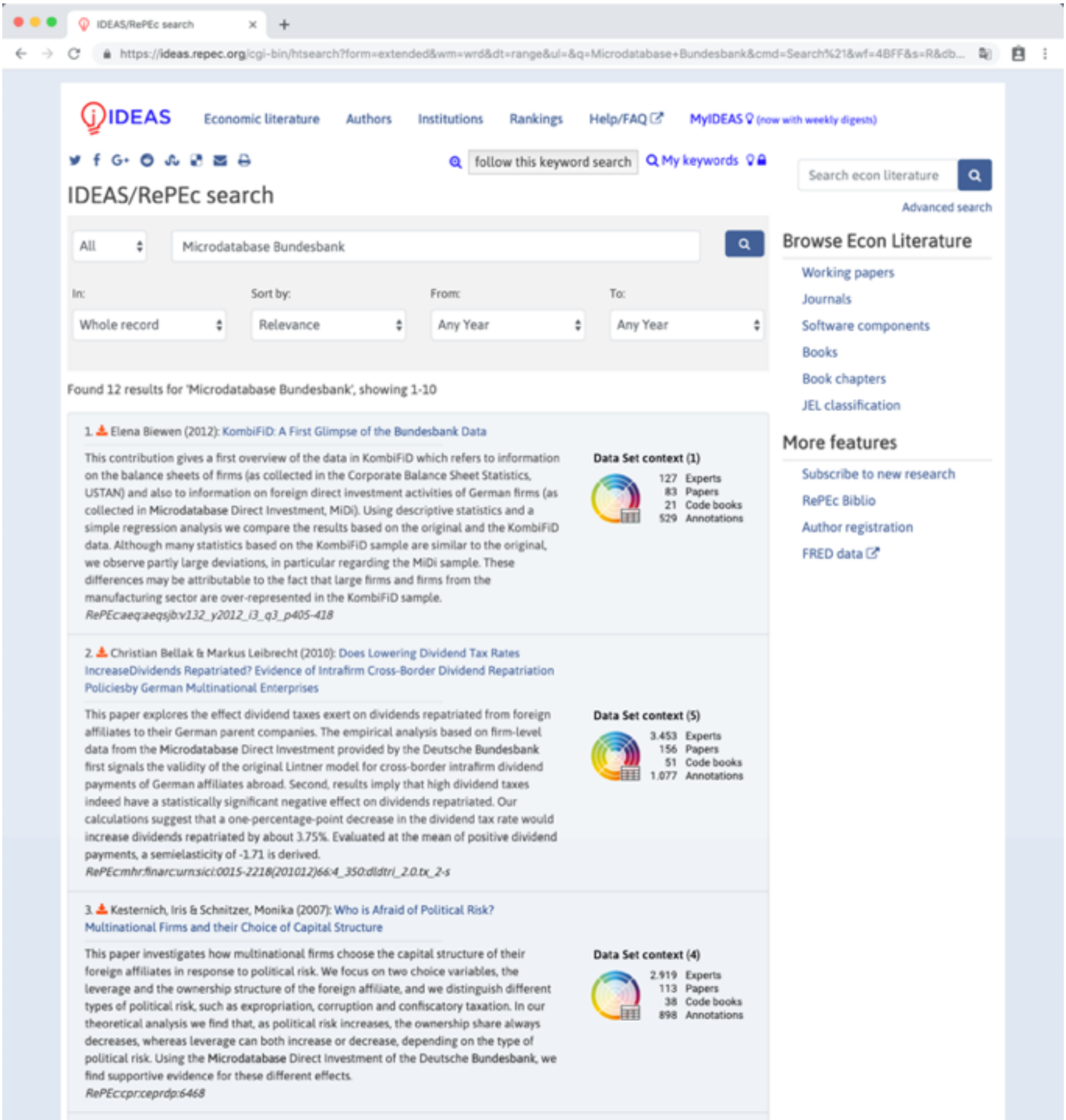
Figure 3.1: Different types of context tracked by Altmetric.com for any research output.
(Reproduced by permission of Altmetric.com)

How datasets are used in research more broadly is another important piece of context that data search engines lack that would significantly enhance discoverability of a dataset and that would consequently increase the value of the data. This is where the Rich Context project can add significant value to a broad research community.

Enhanced context for research data and its impacts could be offered to users in the form of in-app badges and other "signposts" that connect data with its larger context. Such a contextualizing badge could bring together existing data, including not only the number of citations that the dataset has received, but also whether the data has been versioned (through Figshare's repository metadata), discussed online (through Altmetric data), and what kind of tools and insights have been built on top of the data (through rich mining of full-text and citation data available in the ReadCube reference management corpus and in Dimensions).

Correctly developed and accepted by the community, this type of information can make a contribution to solving many of the problems highlighted in this article. If the correct contextual facets can be developed, then recognition would be easier to assign to those who have contributed to the process of creating and maintaining datasets. With greater context around them, datasets become easier to locate, understand and value. This in turn could lead to a broader evaluative environment and more engagement from academics.

Engagement across academia, however, is not uniform. Mechanisms need to be provided to engage data science-focused researchers from whom more details of their tools, scripts and codebooks could be drawn, adding further value to research data. At the same time, engagement tools need to allow data scientists to leverage this information so that it is valuable to them when they are the consumers of search results. These are subtly different use cases from those of standard researchers. By mining ever more open research systems wherein data is being analyzed (e.g. Gigantum, Github, etc), we can start to integrate these other crucial engagement contexts as well.



IDEAS/RePEc search

All: Microdatabase Bundesbank

In: Whole record — Sort by: Relevance — From: Any Year — To: Any Year

Found 12 results for 'Microdatabase Bundesbank', showing 1-10

1. Elena Biewen (2012): KombiFiD: A First Glimpse of the Bundesbank Data

This contribution gives a first overview of the data in KombiFiD which refers to information on the balance sheets of firms (as collected in the Corporate Balance Sheet Statistics, USTAN) and also to information on foreign direct investment activities of German firms (as collected in **Microdatabase** Direct Investment, MiDi). Using descriptive statistics and a simple regression analysis we compare the results based on the original and the KombiFiD data. Although many statistics based on the KombiFiD sample are similar to the original, we observe partly large deviations, in particular regarding the MiDi sample. These differences may be attributable to the fact that large firms and firms from the manufacturing sector are over-represented in the KombiFiD sample.
RePEc:aeq:aeqsjb:v132_y2012_i3_q3_p405-418

Data Set context (1)

127	Experts
83	Papers
21	Code books
529	Annotations

2. Christian Bellak & Markus Leibrecht (2010): Does Lowering Dividend Tax Rates IncreaseDividends Repatriated? Evidence of Intrafirm Cross-Border Dividend Repatriation Policiesby German Multinational Enterprises

This paper explores the effect dividend taxes exert on dividends repatriated from foreign affiliates to their German parent companies. The empirical analysis based on firm-level data from the **Microdatabase** Direct Investment provided by the Deutsche **Bundesbank** first signals the validity of the original Lintner model for cross-border intrafirm dividend payments of German affiliates abroad. Second, results imply that high dividend taxes indeed have a statistically significant negative effect on dividends repatriated. Our calculations suggest that a one-percentage-point decrease in the dividend tax rate would increase dividends repatriated by about 3.75%. Evaluated at the mean of positive dividend payments, a semielasticity of -1.71 is derived.
RePEc:mhr:finarc:urn:sici:0015-2218(201012)66:4_350:dldtri_2.0.tx_2-s

Data Set context (5)

3.453	Experts
156	Papers
51	Code books
1.077	Annotations

3. Kesternich, Iris & Schnitzer, Monika (2007): Who is Afraid of Political Risk? Multinational Firms and their Choice of Capital Structure

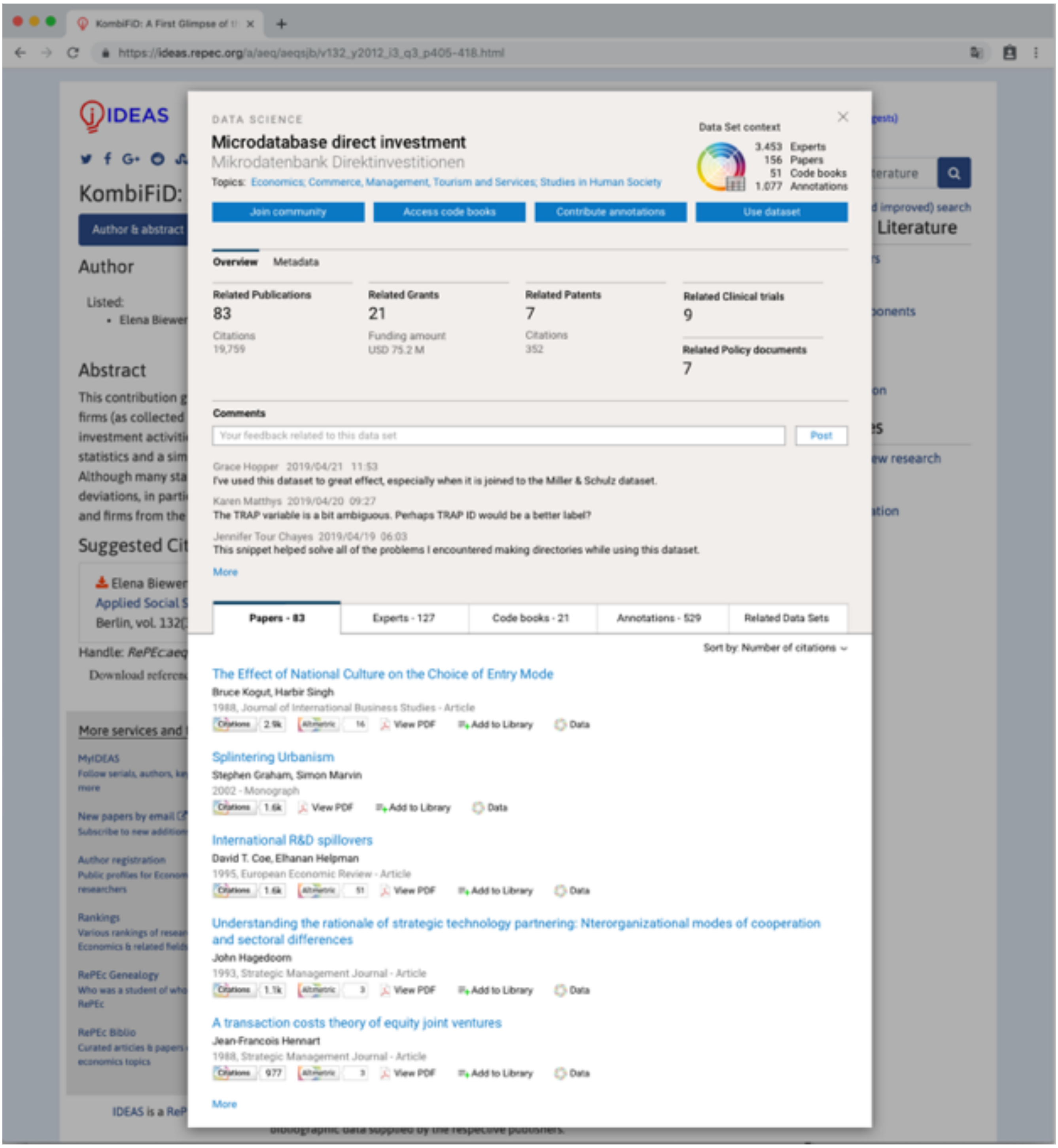
This paper investigates how multinational firms choose the capital structure of their foreign affiliates in response to political risk. We focus on two choice variables, the leverage and the ownership structure of the foreign affiliate, and we distinguish different types of political risk, such as expropriation, corruption and confiscatory taxation. In our theoretical analysis we find that, as political risk increases, the ownership share always decreases, whereas leverage can both increase or decrease, depending on the type of political risk. Using the **Microdatabase** Direct Investment of the Deutsche **Bundesbank**, we find supportive evidence for these different effects.
RePEc:cpr:ceprdp:6468

Data Set context (4)

2.919	Experts
113	Papers
38	Code books
898	Annotations

{width="4.543310367454068in"
height="4.776042213473316in"}

Figure 3.2: Mock-up of a research data badge helping to contextualise a set of search results.



https://ideas.repec.org/a/aeq/aeqsjb/v132_y2012_i3_q3_p405-418.html

DATA SCIENCE

Microdatabase direct investment

Mikrodatenbank Direktinvestitionen

Topics: Economics; Commerce, Management, Tourism and Services; Studies in Human Society

Data Set context

3.453	Experts
156	Papers
51	Code books
1.077	Annotations

Join community | Access code books | Contribute annotations | Use dataset

Overview | Metadata

Related Publications 83 – Citations 19,759

Related Grants 21 – Funding amount USD 75.2 M

Related Patents 7 – Citations 352

Related Clinical trials 9

Related Policy documents 7

Comments

Your feedback related to this data set | Post

Grace Hopper 2019/04/21 11:53
I've used this dataset to great effect, especially when it is joined to the Miller & Schulz dataset.

Karen Matthys 2019/04/20 09:27
The TRAP variable is a bit ambiguous. Perhaps TRAP ID would be a better label?

Jennifer Tour Chayes 2019/04/19 06:03
This snippet helped solve all of the problems I encountered making directories while using this dataset.

More

Papers - 83 | Experts - 127 | Code books - 21 | Annotations - 529 | Related Data Sets

Sort by: Number of citations

The Effect of National Culture on the Choice of Entry Mode
Bruce Kogut, Harbir Singh
1988, Journal of International Business Studies - Article
Citations 2.9k | Altmetric 16 | View PDF | Add to Library | Data

Splintering Urbanism
Stephen Graham, Simon Marvin
2002 - Monograph
Citations 1.6k | View PDF | Add to Library | Data

International R&D spillovers
David T. Coe, Elhanan Helpman
1995, European Economic Review - Article
Citations 1.6k | Altmetric 51 | View PDF | Add to Library | Data

Understanding the rationale of strategic technology partnering: Nterorganizational modes of cooperation and sectoral differences
John Hagedoorn
1993, Strategic Management Journal - Article
Citations 1.1k | Altmetric 3 | View PDF | Add to Library | Data

A transaction costs theory of equity joint ventures
Jean-Francois Hennart
1988, Strategic Management Journal - Article
Citations 977 | Altmetric 3 | View PDF | Add to Library | Data

More

{width="4.163163823272091in"
height="4.505208880139983in"}

Figure 3.3: Mock-up of a research data badge helping to contextualise a specific dataset.

In Figures 3.2 and 3.3, we have visualised some early concepts for how a contextualized research data badge could look. This visualisation is based on insights from the Rich Context project and uses data that could be mined from articles that use a specific dataset. In particular, we suggest four facets of context that both data science-focused researchers and others could find helpful when viewing a dataset:

- **Experts who have made use of the data**, sourced from references made to the dataset in a professional context such as an industry whitepaper or policy document
- **Academics** that **cite the data**, mined from citation of the dataset or ancillary data in the peer reviewed literature
- **End users of the data**, sourced from code book references included in public code repositories
- **Enhancements of the data**, vis-à-vis annotations and comments made on the data in public forums.

In summary, we believe that, if deployed across the many environments in which researchers discover data, the thinking behind the Rich Context project can overcome both the cultural and information-based infrastructure challenges that we highlighted. If these challenges can be overcome by the methods developed, for example in Chapter 13 of this volume, then this will significantly extend the use and discoverability of datasets. The number and variety of datasets in use in academia will certainly expand in the future, and we can only see data becoming even more central to contemporary research efforts. As such, it is critical to invest in robust infrastructures, not only to support the production and sharing of these data, but also to change the culture and evaluative environment around research data. It is only through initiatives such as these that we will be able to solve the vast and complex sociotechnical challenges that face academia today.

References

Allen, L., Scott, J., Brand, A., Hlava, M., Altman, M., 2014. Publishing: Credit where credit is due. Nature 508, 312–313. https://doi.org/10.1038/508312aCERN, 2019. CERN Open Data Portal [WWW Document]. URL http://opendata.cern.ch/search?page=1&size=20&subtype=Collision&type=Dataset (accessed 11.30.19).CISER, 2019. CISER. URL https://ciser.cornell.edu/data/data-archive/ (accessed 11.30.19).DataCite Metadata Working Group, 2016. DataCite Metadata Schema 4.0 [WWW Document]. URL https://support.datacite.org/docs/schema–40 (accessed 11.30.19).Eijnatten, J. van, Pieters, T., Verheul, J., 2013. Big Data for Global History: The Transformative Promise of Digital Humanities. BMGN-LCHR 128, 55. https://doi.org/10.18352/bmgn-lchr.9350Giuliano, F., 2019. Humanités numériques et archives: la longue émergence d’un nouveau paradigme. Documentation et bibliothèques 65, 37. https://doi.org/10.7202/1063788arGrusin, R., 2014. The Dark Side of Digital Humanities: Dispatches from Two Recent MLA Conventions. differences 25, 79–92. https://doi.org/10.1215/10407391–2420009Hook, D.W., Porter, S.J., Herzog, C., 2018. Dimensions: Building Context for Search and Evaluation. Front. Res. Metr. Anal. 3, 23. https://doi.org/10.3389/frma.2018.00023Jisc, 2019. Research Data Management Toolkit | Jisc [WWW Document]. URL https://rdmtoolkit.jisc.ac.uk/plan-and-design/data-management-planning/ (accessed 11.30.19).Jordan, K., Weller, M., 2018. Academics and Social Networking Sites: Benefits, Problems and Tensions in Professional Engagement with Online Networking. Journal of Interactive Media in Education 2018, 1. https://doi.org/10.5334/jime.448Kamp, K., Herbell, K., Magginis, W.H., Berry, D., Given, B., 2019. Facebook Recruitment and the Protection of Human Subjects. West J Nurs Res 41, 1270–1281. https://doi.org/10.1177/0193945919828108Leurs, K., Shepherd, T., 2017. 15. Datafication & Discrimination, in: The Datafied SocietyStudying Culture through Data. Amsterdam University Press, Amsterdam. https://doi.org/10.1515/9789048531011–018Li, N., Li, T., Venkatasubramanian, S., 2007. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, in: 2007 IEEE 23rd International Conference on Data Engineering. Presented at the 2007 IEEE 23rd International Conference on Data Engineering, IEEE, Istanbul, pp. 106–115. https://doi.org/10.1109/ICDE.2007.367856Linek, S.B., Fecher, B., Friesike, S., Hebing, M., 2017. Data sharing as social dilemma: Influence of the researcher's personality. PLoS ONE 12, e0183216. https://doi.org/10.1371/journal.pone.0183216NNLM, 2019. Data Management Plan | NNLM [WWW Document]. URL https://nnlm.gov/data/data-management-plan (accessed 11.30.19).Siddle, J., 2014. I Know Where You Were Last Summer: London's public bike data is telling everyone where you've been. I Know Where You Were Last Summer. URL https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html (accessed 11.30.19).Stark, L., 2018. Algorithmic psychometrics and the scalable subject. Soc Stud Sci 48, 204–231. https://doi.org/10.1177/0306312718772094Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., Hayashi, K., O'Donnell, D., Digital Science, Hook, D., 2016. The State of Open Data Report. Digital Science. https://doi.org/10.6084/M9.FIGSHARE.4036398.V1

Biographical information

Christian Herzog is CEO of Dimensions and Chief Portfolio Officer at Digital Science. A medical doctor by training, Christian also studied economics and started in 2005 Collexis, a software company focused on text-mining based software applications for the research space. In 2010, Collexis was acquired by Elsevier where Christian spent the following two years as the VP for Product Management SciVal. in 2013, Christian and his co-founders started ÜberResesarch as part of Digital Science which led to the launch of Dimensions as a large-scale research information infrastructure in 2018.

Daniel Hook is CEO of Digital Science. He co-founded Symplectic while studying for his PhD in theoretical physics at Imperial College London in 2003. Symplectic became one of Digital Science's first investments in 2010. Daniel continues to be an active researcher and holds visiting academic positions at Imperial College London and at Washington University in St Louis. He has written more than 30 academic papers and has co-authored a book on Quantum Theory. Daniel is a Fellow of the Institute of Physics, a Policy Fellow at CSaP in Cambridge and serves on the ORCID board as its treasurer.

Mark Hahnel is the CEO and founder of Figshare, which he created whilst completing his PhD in stem cell biology at Imperial College London. Figshare provides research data infrastructure for institutions, publishers and funders globally. Mark is passionate about open science and its potential to revolutionize the research and has led the community in the development of research data infrastructure. Mark sits on the DataCite board, the DOAJ advisory board, the judging panel for the National Institutes of Health (NIH), Wellcome Trust Open Science prize and acted as an advisor for SpringerNature's masterclasses.

Stacy Konkiel is the Director of Research Relations at Dimensions and Altmetric. Stacy's research interests include incentives systems in academia and informetrics, and she has written and presented widely about altmetrics, Open Science, and research data services. Previously, Stacy worked with teams at Impactstory, Indiana University & PLOS. You can learn more about Stacy at stacykonkiel.org.

Duane Williams is VP of US Government at Digital Science. Duane earned his doctorate in theoretical chemistry from the Quantum Theory Project at University of Florida. His current work focuses on improving strategic research investment decisions through new data sets, new tools and metrics to gain insight into the global research landscape. Prior to joining Digital Science, he served as Senior Scientific Analyst and

Project Manager for the IP and Science division of Thomson Reuters (now Clarivate Analytics). There he designed and led custom analyses and software development to facilitate data-driven objective assessments of research programs.

Chapter 4 - Metadata for Social Science Datasets

Metadata for Administrative and Social Science Data

Robert B Allen

[0000–0002–4059–2587]

rba\@boballen.info

Data are valuable but finding the right data is often difficult. This chapter reviews current approaches to metadata about numeric data and considers approaches that may facilitate the identification of relevant data. In addition, the chapter reviews how metadata support repositories, portals, and services. There are many emerging metadata standards, but they are applied unevenly so that there is no comprehensive approach. There has been greater emphasis on structural issues than on semantic descriptions.

INTRODUCTION

Evidence-based policy needs relevant data (Commission on Evidence-Based Policy, 2018; Lane, 2016). Such data is often difficult to find and use. The FAIR Open Access guidelines suggest that, ideally, data should be Findable, Accessible, Interoperable, and Reusable (FAIR).[1] Broad and consistent metadata can support these needs. Metadata and other knowledge structures could also supplement and ultimately even replace text.

This chapter surveys the state of the art of metadata for numeric datasets, focusing on metadata for administrative and social science records. Administrative records describe details about the state of the world as collected by organizations or agencies. They include governmental, hospital, educational, and business records. By comparison, social science data generally is collected for the purpose of developing or applying theory.

We start by considering data and datasets (Section 2) and basic principles of metadata and their application to datasets (Section 3). Modern metadata is often implemented with Resource Description Framework (RDF) linked data (Section 4). Section 5 introduces ontologies and other semantic approaches. We then move to applications which use metadata. Section 6 examines repositories that hold and distribute collections of datasets. Section 7 describes services and techniques associated with repositories and Section 8 briefly describes the computing infrastructure for repositories.

DATA ELEMENTS AND DATASETS

While data may be incorporated in text, image, or video, here we focus on numeric observations recorded and maintained in machine-readable form. Individual observations are rarely used in isolation. Rather, they are typically collected into datasets.

A dataset is defined in the W3C-DCAT (W3C - Data Catalog Vocabulary)[13] as "a collection of data, published or curated by a single agent"[14] such as a statistical agency. There are many different types of datasets; they differ in their structure, their source, and their use. A given data element may appear in many different datasets and may be numerically combined with other data to form derived data elements which then appear in still other datasets. In some cases, they are single vectors of data; in other cases, they comprise all the data associated with one study or across a group of related datasets. Reference datasets are generally collected and archived because they are of enduring value and can be used for answering many different types of questions. Other datasets, such as an individual's medical records, are associated with a relatively narrow set of applications.

There is wide variability in the organization and contents of datasets, as well as in the extent to which datasets are validated and curated. Potentially with frameworks such as the SDMX (Statistical Data and Metadata eXchange) Guidelines for the Design of Data Structure Definitions,[15] concise structured descriptions can be developed for how data elements are combined to form datasets.

METADATA SCHEMAS AND CATALOGS

Many datasets are available; the DataCite repository alone contains over five million datasets. Metadata can support users in finding datasets and enable users to know what is in them. Metadata are short descriptors which refer to a digital object. However, there is tremendous variability in the types of metadata and how they are applied. One categorization of metadata identifies structural (or technical), administrative, and descriptive metadata (Riley, 2017). Structural metadata includes the organization of the files. Administrative metadata describes the permissions, rights, preservation and usage relating to the data.[16] Descriptive metadata covers the contents.

A metadata element describes an attribute of a digital object. The simplest metadata (e.g., a Digital Object Identifier (DOI) or ORCiD[17]) identifies the digital object or its creator.[18] Metadata elements are generally part of a schema, or frame. DCAT[19] is a schema standard for datasets that is used by many repositories such as data.gov. Other structured frameworks for datasets include the DataCite[20] metadata schema and the Inter-university Consortium for Political and Social Research Data Documentation Initiative (ICPSR DDI, see Section 6.1). ISO 19115–1:2014 establishes a schema for describing geographic information and services. [21]

The schema specifications provide a flexible framework. For instance, DCAT allows the inclusion of metadata elements drawn from domain schema and ontologies. Some of these domain schemas are widely used resources which DCAT refers to as assets. Figure 1 shows a fragment of properties (i.e., metadata elements) from an implementation of the Schema.org[22] dataset schema to describe gross domestic product.

Figure 4.1: Fragment of GDP properties described by the Schema.org dataset schema.[23]

Metadata terms for an application are often assembled into namespaces from different metadata schemas. Metadata Application Profiles[24] provide constraints on the types of entities that can be included in the metadata for a given application. Moreover, application profiles can be used to validate standards. For instance, the DCAT Application Profile for data portals in Europe (DCAT-AP) supports the integration of data drawn from repositories in different jurisdictions in the EU.[25]

A collection of dataset schema,[26] such as all the datasets in a repository, forms a catalog. For data streams, there needs to be continuity but also the ability to update the records. In some cases, there may be relatively infrequent periodic updates. These could be given version numbers rather than an entirely new DOI.[27] However, collections of highly dynamic data streams present challenges; most of the data stay the same but some of the data and/or metadata (e.g., number of records) change.

LINKED DATA

RDF (Resource Description Framework) extends XML by requiring triples which assert a relationship (property) between two identifiers: “identifier – property - identifier”. RDF Schema (RDFS) extends RDF by supporting subclass relationships. A graph is formed by linking triples.

Hierarchical classification systems are another knowledge structure with a long history. Indeed, Schema.org is based around a hierarchical ontology. Simple classification relationships are handled by the Simple Knowledge Organization System (SKOS). SKOS represents the hierarchical structure of traditional thesauri with RDFS. Collections of data organized by SKOS are often described as “linked data”.

Depending on the rigor with which they are developed, these collections can support limited logical inference. Many administrative and social-science-related thesauri, such as EDGAR and those of the World Bank and the OECD, have now been implemented with SKOS. A knowledgebase is, primarily, a SKOS graph that links real-world entities. For example, Wikidata[28] is an effort to develop a knowledgebase based on structured data from Wikimedia projects, and VIVO[29] is a knowledge graph of scholarship.

But there are also many stand-alone classification schemes. The Extended Knowledge Organization System (XKOS)[30] was developed to allow classification systems to be incorporated into a SKOS framework.

RICHER SEMANTICS

Ontologies provide a coherent set of relationships between entities which cover a given domain. Well-constructed ontologies can support logical inference. Some vocabularies such as Dublin Core, which is implemented in RDF, are said to have an ontology, but they are limited because relationships among the terms are not specified. FOAF (Friend of a Friend) provides a somewhat richer ontology which includes attributes associated with people. Still more extensive ontologies often use OWL (Web Ontology Language) which can support stronger logical inference than RDFS.

One way to coordinate across terms is an upper ontology. Upper ontologies provide top-down structures for the types of entities allowed in domain and application ontologies. One of the best-known upper ontologies is the Basic Formal Ontology (BFO) (Arp, Smith, & Spear, 2015), which is a realist, Aristotelian approach. At the top-level, BFO distinguishes between Continuants (endurants) and Occurrents (perdurants) and also between Universals and Particulars (instances). Many biomedical ontologies based on BFO are collected in the Open Biomedical Ontology (OBO) Foundry.[31]

There are fewer rich ontologies dealing with social science content than for natural science. Social ontology, that is, developing rigorous definitions for social terms, is often a challenge. It is difficult to define precisely what is a family, a crime, or money. In most cases, an operational or approximate definition may suffice when formal definitions are difficult. However, those operational definitions often do not interoperate well across studies.

DATA REPOSITORIES AND COLLECTIONS OF DATASETS

A data repository holds datasets and related digital objects. Ideally, it contains a stable collection selected according to a collection policy. It is organized by metadata and knowledge structures. It provides access to the datasets and typically supports search.

The Inter-University Consortium for Political and Social Research (ICPSR)

ICPSR[32] is a major repository of public-use social science and administrative datasets derived mostly from questionnaires and surveys. We go into depth about it here because the ICPSR Data Documentation Initiative (DDI)[33] (e.g., Vardigan, Heus, & Thomas, 2009) is especially well-crafted.[34] The DDI codebook saves the exact wording of all the questions and ICPSR provides an index of all variable names. DDI-Lifecycle is an extension that describes the broader context in which the survey was administered as well as the details about the preservation of the file (see Section 7.1). **DDI uses XKOS to provide linked data.** Figure 2 shows the ICPSR DDI metadata schema.

Figure 4.2: ICPSR DDI metadata elements.

Version

Study Title

Alternate Title

PIs & Affiliation

Funding Agencies

Summary

Subject Terms

Geographic Coverage Areas

Geographic Representation

Study Time Periods and Time Frames

Collection Notes

Study Purpose

Study Design

Description of Variables

Sampling: Sampling Procedure, Sampling Unit, Sampling Notes

Oversampled Group

Time Method

Data Source Type

Mode of Collection

Weight

Response Rates

Scales

Analysis Unit

Unit of Observation

Smallest Geographic Unit

Data Format

Restrictions

Version History

The ICPSR metadata elements incorporate aspects of the implementation and design of research studies. However, many of the ICPSR metadata elements are not independent; potentially, they could be interlinked with terms such as organizations, locations, individuals, and research designs from other knowledgebases. Moreover, they could be linked with higher-level workflows and mechanisms (see Sections 7.1, 7.5).

Additional Examples of Repositories

Statistical data collection is a core function of government. Such collections often emphasize social data such as employment, criminal justice, and public health. They also include related indicators such as agricultural and industrial output and housing. Most countries have national statistical agencies such as Statistics New Zealand, and the Korean Social Science Data Archive (KOSSDA). European datasets are maintained in the Consortium of European Social Science Data Archives (CESSDA)[35] and the European Social Survey.[36] Australia has a broad data management initiative, ANDS.[37] Many U.S. federal governmental datasets are collected at data.gov. In addition, there are many other social survey repositories[38] and many U.S. states and cities have online statistics sites at varying levels of sophistication.

There are also many non-governmental and inter-governmental agencies such as the OECD, the World Bank, and the United Nations that manage datasets. Similarly, there are very large datasets from medical research such as from clinical trials and from clinical practice including Electronic Health Records (EHRs).

Many datasets are produced, curated, and used in the natural sciences such as astronomy and geosciences. Some of these datasets have highly automated data collection, elaborate archives, and established curation methods. Many repositories contain multiple datasets for which access is supported with portals or data cubes (see Section 7.4). For instance, massive amounts of geophysical data and related text documents are collected in the EarthCube[39] portal. The science.gov portal is established by the U.S. Office of Science Technology and Policy. NASA supports approximately 25 different data portals. Each satellite in the Earth Observation System (EOS) may provide hundreds of streams of data,[40] with much common metadata. Likewise, there are massive genomics and proteomics datasets which are accessible via portals such as UniProt[41] and the Protein Data Bank[42] along with suites of tools for exploring them.

Repository Registries

There are a lot of different repositories, so it is useful to have a registry with a standard schema structure for describing them. The Registry of Research Data Repositories[43], which is operated by DataCite, links to more than 2000 repositories each of which holds many datasets. Each of those repositories is described by the re3data.org schema (Rücknagel, Vierkant, et al., 2015).

Ecosystems of Texts and Datasets

Datasets are often associated with text reports, whether they describe the development of the datasets or their use. Ultimately, we would like to be able to move seamlessly from datasets to texts and other related materials. However, as demonstrated by several of the papers in this volume, it is often difficult to extract details about datasets from legacy publications.

Text associated with a dataset may be used to support searching for it. Indeed, Google Dataset Search uses texts marked up with Schema.org JSON-LD (JavaScript Object Notation for Linked Data) micro-data to generate an index.

Going forward, great value can be achieved by persuading editors and authors to clearly cite and deposit datasets. In some cases, a separate data editor may be appointed. The Dryad Digital Repository[44] captures datasets from scholarly publications. It requires the deposit of data associated with scholarly papers accepted for publication. Such datasets are most often used to validate the conclusions of a research publication, but they may also be used more broadly.

Research datasets may be given DOIs[45] and cited in much the same way that research reports are cited. Formal citations can support tracing the origins of data used in analyses and help to acknowledge the work of the creators of the datasets.

Information Institutions and Organizations

The Open Archival Information System (OAIS) provides a reference model for the management of archives (Lee, 2010). A key part of the model is the inclusion of preservation planning and the requirement for stable administration over time. These attributes are part of all information institutions. Libraries, archives, and museums have formal collection management strategies, metrics, and policies.

In addition to traditional information institutions, there are now many other players. CrossRef[46] and DataCite are DOI registration agencies. CrossRef is a portal to metadata for scholarly articles, while DataCite provides metadata for digital objects associated with research. Schema.org's primary mission is to provide a structure that improves indexing by search engine companies. Still other organizations such as

HL7[47] and KEGG[48] manage controlled vocabularies and frameworks. These organizations are increasingly adopting best practices similar to those of traditional information organizations.

REPOSITORY SERVICES

Administrative Metadata and Related Services

Administrative metadata is one of the three broad categories of metadata. Administrative metadata describes the permissions, rights, preservation, and usage of the data. While the focus of a traditional library is to support access and the focus of an archive is to ensure stability and quality, increasingly, digital repositories must address both access and preservation.

Preservation and Trusted Datasets: Although data storage prices are declining dramatically, the cost of maintaining a trusted repository remains substantial and we cannot save everything. These challenges are familiar from traditional archives; selection policies typical in archives could help in controlling the many poorly documented datasets in some repositories. Yet, prioritization of what to select is difficult (Whyte & Wilson, 2010)[49].

Lost data is often irreplaceable. Even if the data is not entirely lost, users need confidence that the validity of stored data has not been compromised. Indeed, some data may become the target of malicious attacks. Trust is a result of both technology and organizational procedures. Technology may include hash-based encoding of data. CLOCKSS[50] is a distributed hash system for web-based scholarly literature. Blockchains provide hashed records of transactions and can be applied to data records.

The OAIS framework has been incorporated into the ICPSR DDI-Lifecycle model. The integrated Rule-Oriented Data System (iRODS)[51] is a policy-based archival management system[52] developed for large data stores. It implements a service-oriented architecture (SOA) to support best practices established by archivists. Further, audits, such as by the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA),[53] may be conducted to assess how well repositories implement trustworthy procedures.

Preservation and provenance metadata schemes such as PREMIS[54] and PROV-O[55] are state-based ontologies that include entities such as actors, events, and digital objects. They record the history of transitions (e.g., changes in format) for digital objects.

Rights Metadata: For some data, there are many advantages to open publication. The rights for that data can be specified with a Creative Commons License. For other data, there can be strong justifications for limited access, such as privacy and economic factors.

For example, although survey results are generally aggregated across individuals, individual-level data is sometimes very useful. Some repositories of survey data include micro-data, that is data for the responses that individuals gave to survey questions.[56] However, analysis of such micro-data raises privacy concerns and needs to be carefully managed; access should be limited to qualified researchers. Repositories of individual health records raise similar privacy concerns.

Usage Statistics: The number of visits and downloads for a dataset can give an indication to later users about the likely value of a given dataset. Such usage data are helpful for the managers and funders of repositories to evaluate their service. Citations are indicators for how a dataset is being used and its relationship to other work.

Analysis Platforms and Decision Support Systems

There is an increasingly rich set of analytic tools. Some of the earliest tools were statistical packages such as SPSS, R, SAS, and STATA. These were gradually enhanced with data visualization and other analytic software. The current generation of tools such as Jupyter[57], RSpace, and eLab notebooks (ELN) integrate annotations, workflows, raw data, data analysis, and annotations into one environment.

Virtual research environments (VREs) are typically organized by research communities to coordinate datasets with search and analytic tools. For instance, the Virtual Astronomy Observatory (VAO) uses Jupyter to provide users with a robust research environment. WissKI[58] is a platform for coordinating digital humanities datasets which are based on Drupal. Decision Support Systems (DSS) are generally focused on finding optimal solutions in a parameter space. They often draw on data warehouses though recently they have begun to incorporate feeds from unstructured data (e.g., web searches).

Most repositories support search on metadata terms. In addition, some repositories have developed their own powerful data exploration tools such as ICPSR Colectica[59] for DDI and the GSS Data Explorer[60]. The Amundsen data discovery and metadata engine[61] uses metadata elements to provide a table explorer. Potentially, interactive visualization tools such as TableLens (Rao & Card, 1994) could also be employed.

Metadata Development, Standardization, and Management

Metadata, whether for texts or datasets, needs to be complete, consistent, standardized, machine processable, and timely (Park, 2009). Metadata registries provide clear definitions and promote standardization (ISO/IEC 11179). For instance, the Marine Metadata Interoperability Ontology Registry and Repository[62] records usage of different metadata terms. A registry may interoperate with editing tools for developers (Goncalves, O'Connor, et al., 2019). These tools may suggest candidate metadata terms. One of the keys to the development of good metadata is the involvement of a community that cares about the results.

Data Cubes, Data Warehouses, and Data Exchanges

An organization such as a large business often has many different databases. The data in the databases will likely have different formats and definitions and can be organized in a multidimensional cube. Some of cube's cells may be well-populated with data that appears across many of the databases, but there will also be sparsely populated regions and cells. Online Analytical Processing (OLAP) users can generate different views of the data by drilling-down, rolling-up, and slicing-and-dicing across cells. To facilitate retrieval, there can be a rich pre-coordinated index for common queries. Other queries can be implemented with slower methods such as hashing or B-trees.

While many organizations now have integrated enterprise data management systems, data cubes are still useful for warehousing data and for exchanging it across organizations. For instance, the W3C Data Cube[63] standard is applied in inter-organizational projects such as EarthCube[64]. *SDMX*[65] *enables data exchange among statistical agencies in the EU.*

Production Workflows, Research Workflows, and Research Objects

Entities change over time, yet many knowledge representation frameworks do not model change. To represent change, models need to represent transitions, processes, and other sequential activities. Such modeling is closer to state machines, Petri Nets, process ontologies, the Unified Modeling Language (UML) or even programming languages than to traditional knowledge representation.

One way to document a research project is by saving files developed during the study (Borycz & Carroll, 2018). Data files (e.g., Excel files) are just one type of artifact from a research program; other research objects include workflows. Workflows are a natural fit for describing research methods and analyses (Austin, et al., 2017). The Taverna[66] workflow tool has been used for the MyExperiment[67] project. It provides a framework for capturing and posting Taverna and other types of research workflows and incorporates simple ontologies such as FOAF. Workflows can also be used to specify and document statistical analyses; several of the analysis platforms described in Section 7.2 support them. Sequential activities in the management of repositories are often tracked with workflows. For instance, the Generic Statistical Information Model (GSIM)[68] specifies workflows for the production of datasets by statistical agencies.

Semantic Modeling and Direct Representation

Semantic models attempt to represent entities. They could support unified descriptions of functionality, transitions of complex continuants, and sequential activities (Allen, 2018). Changes in semantic models are a form of qualitative simulation. While traditional knowledge representation is usually implemented with ontologies, models which allow transitions are more like programming languages. Although semantic modeling might be implemented by process ontologies, we have focused on the use of an object-oriented programming language which supports threads to allow parallel concurrent event streams and potentially to develop a “unified temporal map”. Such semantic simulations may be useful for modeling historical events. For instance, a community described in a newspaper may be cast into a “community model”. These go beyond social ontology to model social mechanisms (Ylikoski, in press).

In addition, Allen (2015, 2018) has proposed rich semantic modeling of entire research reports and datasets. Structured evidence and argumentation about claims might then be applied for the evaluation of the models. Ultimately, such “direct representation” may replace text as the primary representation for research and scholarship.

INFRASTRUCTURE

Repository Servers

Semantic representations may be implemented with triplestores. Triplestores facilitate logical inference, but retrieval may be more efficient with relational databases. Many metadata catalogs are implemented with relational databases. Thus, they use SQL and are often characterized by UML Class Diagrams. Information models (e.g., National Information Exchange Model (NIEM)[69]) which could be used for metadata registries may be implemented as data dictionaries.

Some repositories are federated with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[70]. This allows the “harvesting” of metadata from separate repositories. OAI-PMH is increasingly used as an API to allow external users to query and interact with the federated set of metadata.

Cloud Computing

We are well into the era of cloud computing (Foster & Gannon, 2017), allowing flexible allocation of computing, networking and storage resources, which facilitates Software as a Service (SaaS). The compatibility of the versions of software packages needed for data management is often a challenge. Containers, such as those from Docker, allow compatible versions of software to be assembled and run on a virtual computer. A cloud-based virtual machine can hold datasets, workflows, and the programs used to analyze the data, which can be a complete digital preservation package.

Highly networked data centers facilitate the Internet of Things (IoT) which generates massive and dynamic data. Increasingly, cloud computing is supporting edge computing and append-only stores which can capture streaming data. These technologies will provide the foundation of smart cities and have implications for the kinds of questions we may ask about social behavior.

CONCLUSION

Many datasets, especially legacy datasets, are difficult to find and access. Some of the biggest issues for the retrieval of datasets concern information organization, which helps to provide context. Metadata supports the discovery and access to datasets.

More attention to metadata would also further support evidence-based policy. We need richer, more systematic, and more interoperable metadata standards. We need to improve the metadata associated with existing datasets. And we need to aggressively upgrade the application of high-quality metadata and knowledge organization systems to datasets as they are created.

ACKNOWLEDGMENTS {#acknowledgments .ListParagraph}

Julia Lane and members of NYU’s Center for Urban Science and Progress provided useful advice and comments.

REFERENCES {#references .ListParagraph}

Allen, R.B. (2015) Repositories with direct representation, *Networked knowledge representation systems,* arXiv: 1512.09070

Allen, R.B. (2018) *Issues for using semantic modeling to represent mechanisms*, arXiv:1812.11431

Allen, R.B., & Kim, YH. (2017/2018) Semantic modeling with foundries, arXiv:1801.00725

Arp, R., Smith, B., & Spear, A.D. (2015) *Building ontologies with basic formal ontology*, MIT Press, Cambridge. MA.

Austin, C.C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V.K., Murphy, F., Nurnberger, A., et al. (2017). Key components of data publishing: Using current best practices to develop a reference model for data publishing. *International Journal on Digital Libraries, 18*(2) 77–92, doi:10.1007/s00799–016–0178–2

Borycz, J., & Carroll, B. (2018) Managing digital research objects in an expanding science ecosystem: 2017 conference summary. *Data Science Journal*, 17, doi: http://doi.org/10.5334/dsj–2018–016

Commission on Evidence-Based Policymaking (2018) *The Promise of Evidence-Based Policymaking*, https://www.cep.gov/cep-final-report.html

Foster, I., & Gannon, D.B. (2017) *Cloud computing for science and engineering*, MIT Press, Cambridge, MA.

Gonçalves, R.S., O'Connor, M.J., Martínez-Romero, M., Egyedi, A.L., Willrett, D., Graybeal, J., & Musen, M.A. (2019) *The CEDAR workbench: an ontology-assisted environment for authoring metadata that describe scientific experiments*. arXiv: 1905.06480

InterPARES2 Project (2008) A framework of principles for the development of policies, strategies and standards for the long-term preservation of digital records,

Lane, J. (2016) Big data for public policy: The quadruple helix, *Journal of Policy Analysis and Management, 35*(3), doi: **10.1002/pam.21921**

Lee, C.A. (2010) Open Archival Information System (OAIS) reference model. *Encyclopedia of library and information sciences* (3rd Edition). CRC Press, doi: 10.1081/E-ELIS3–120044377

Park, J-R., (2009) Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly, 47*, 213–228, 2009, doi: 10.1080/01639370902737240

Rao, R., & Card, S.K. (1994) The Table Lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information, *ACM SIGCHI*, 318–322, doi: 10.1145/191666.191776

Riley, J. (2004) *Understanding metadata: What is metadata, and What is it for?: A primer*, NISO Press, Bethesda, MD. ISBN: 978–1–937522–72–8

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D. et al. (2015) *Metadata schema for the description of research data repositories: version 3.0* (29), doi: 10.2312/re3.008

Vardigan, M., Heus,P., & Thomas, W. (2008) Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation.* ***3*****(1). doi:**
10.2218/ijdc.v3i1.45

Whyte, A., & Wilson, A. (2010) How to appraise and select research data for curation. *DCC How-to Guides. Edinburgh*: Digital Curation Centre. http://www.dcc.ac.uk/resources/how-guides

Wilkinson, M.D.,
Dumontier, M.,
Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak. A., et al. (2016) The FAIR guiding principles for scientific data management and stewardship, *Scientific Data, 3*, 160018. doi: 10.1038/sdata.2016.18

Ylikoski, P. (to appear) Social mechanisms. *The Routledge handbook of mechanisms and mechanical philosophy*, Routledge, edited by S. Glennan and P. Illari

Chapter 5 - Competion Design

By Andrew Gordon, Ekaterina Levitskaya, and Jonathan Morgan - New York University

Table of Contents

Introduction

The rich context competition was designed to inspire computer scientists to automate the discovery of research datasets and the associated research methods and fields in social science research publications. Participants were asked to use any combination of machine learning and data analysis 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 competition had the potential to draw on existing work. The IARPA FUSE program had funded research teams to develop automated methods that would identify technical emergence using information found in published scientific, technical, and patent literature(*1*, *2*), and resulted in recommendation systems like meta.com(*3*). Google Dataset Search had developed search technologies that would find datasets in data repositories across the Web(*4*) Academic social network sites like ResearchGate and Academia.edu had developed platforms whereby researchers provided feedback about their scientific activities(*5*) And there is a well established tradition of competitions in computer science, particularly natural language processing(*6*).

Context

Social scientists might define rich context as a dataset search and discovery process: what does the data **measure**, what **research** has been done by which **researchers**, with what **code**, and with what **results**. Computer scientists might define rich context as knowledge graph representation and recommender systems. Others might define rich context as promoting datasets to be a first class entity. But the core idea of the competition was to incentivize computer scientists to build automated tools that find datasets mentioned in scientific publications and build an associated community of interest. The results could then be used to recommend datasets to empirical researchers and encourage researchers to provide feedback about the value of the recommendations.

The innovation literature provided some guidance. At a high level, most systematic incentives for innovation can be classified as one of two types: up-front support for research ("push programs") or commitments to reward successful results ("pull incentives")(*7*), with a given incentive evaluated on its balance between positive and negative outcomes. The patent system of protecting intellectual property for a period of time, for example, is an incentive that balances the benefit to the creator of time-limited exclusive use of a patented innovation with the cost of restriction on broader use(*8*).

Prizes are another common pull incentive, offering direct reward for an innovation that arises from competition among innovators. Innovation prizes offer an immediate benefit that can be a powerful incentive for development and diffusion of innovations, but the design of the contest that awards them is important to maximizing innovation benefits, and effective evaluation is difficult. For prizes to encourage innovations that are of high quality, desirable, and more production-ready, the contests that offer them need to be designed carefully to include additional evaluation requirements or incentives, with the benefits to participants carefully balanced so that the rewards make the additional requirements worth their cost(*9*, *10*) .

The literature informing the development of a community of practice in a domain of work where knowledge is cumulative emphasizes the advantages. Successful communities can develop knowledge-sharing and dissemination mechanisms, common norms of sharing and cooperation, and broad agreement on technical paradigms and jargon(*11*). As open source software communities show, however, they must be carefully incentivized and nurtured to grow participation(*12*) and managed well to maintain resources and quality of output over time(*13*).

Specific Challenges

There were a number of challenges associated with developing a natural language processing (NLP) competition applied academic publications. Access to scientific publications is typically limited. In addition, there are no existing annotated data sets or standards for annotations and existing solutions are not easily reusable. A similar project focused on text analysis for clinical studies reported that NLP research teams do not traditionally collaborate closely, and models and systems that result tend to not be designed or implemented to be easy to use or to scale up for production use(*14*). However, in the NLP domain, Ian Soboroff at the National Institutes of Standards and Technology (NIST) has developed a series of competition patterns designed to inspire disparate groups of researchers to help to carry out information tasks against text data. These include basic competitions where data is provided to groups and they are allowed to train and then submit a number of runs of their models against a subset of evaluation data (*6*). More elaborate competitions include ones organized around an "incident", where groups are given training data and model specifications and allowed to train a model, game out an incident where an event occurs in a previously unseen language and then they have to quickly adapt their model to the new language and submit results(*15*).

We also wanted to encourage researchers to develop a generalized model to identify datasets that was not overly dependent on the use of formal titles of data sets, because many research datasets do not have such formal titles. Thus the problem was much more complicated than a named entity recognition problem, because competitors needed to be able to characterize the language of discussing and using data to recognize where data is discussed in a particular article and then identify which data sets.

This chapter describes the implementation of the competition, particularly focusing on the lessons learned.

Competition Design

The goal of the competition was to use any combination of machine learning and data analysis methods to identify the datasets mentioned in a corpus of social science publications and infer both scientific methods used in the analysis and the publication's research fields[1].

The competition had two phases.

In the first phase, participating teams were provided with a listing of datasets and a labeled corpus of 5,000 publications with an additional dev fold of 100 publications. Each publication was labeled to indicate which of the datasets from the master list were referenced within and what specific text was used to refer to each dataset. The teams used this data to train and tune algorithms to detect mentions of data in publication text and, when a data set in our list is mentioned, tie each mention to the appropriate data set. A separate corpus of 5,000 labeled publications was held back to serve as an evaluation corpus. Each team was allowed up to 2 test runs against this evaluation corpus before final submission. The final models of each group were run against this holdout corpus and the results were used to evaluate submissions, along with a random qualitative review of the mentions, methods, and fields detected by the team's model. Submissions were primarily scored on the accuracy of techniques, the quality of documentation and code, the efficiency of the algorithm, and the quality and novelty of the methods and research fields inferred for each of the publications.

Four finalist teams were selected to participate in the second phase, the teams from: Allen Institute for Artificial Intelligence, United States; GESIS at the University of Mannheim, Germany; Paderborn University, Germany; and KAIST in South Korea.

In the second phase, finalists were provided with a new training corpus of 5000 unlabeled publications and asked to discover which of the datasets from the first phase's data catalog were used in each publication, as well as infer associated research methods and fields. As in the first phase, teams were scored on the accuracy of their techniques, the quality of their documentation and code, the efficiency of their algorithm, and the quality and novelty of the methods and research fields inferred for each of the publications.

At the end of each phase, competing teams packaged their models into a docker container using a model packaging framework designed and built for the competition by NYU, and the containers were installed on AWS servers and run by the competition organizers against the holdout to generate predictions that were used to evaluate the teams.

Data

For training and evaluation data, our goal was to lay the foundations for developing a “gold standard corpus” (GSC) of academic populations tagged with the semantic context within which datasets are mentioned used in analysis. A GSC corpus is one that is manually tagged and reviewed for quality, usually for a particular domain and task. Creating one is time-consuming and expensive(*16*) because it involves selecting a corpus to annotate, then implementing a manual annotation and review scheme[13].

While our goal was not to make a GSC, we used our data creation to begin to assess data needed for high quality data detection models and to test potential methods for creating a GSC. To create our competition training and evaluation data, we started with data set citation data from the ICPSR data catalog ([[https://www.icpsr.umich.edu/icpsrweb/]{.underline}](https://www.icpsr.umich.edu/icpsrweb/)), then used methods that originated in quantitative content analysis of communication artifacts(*17*) combined with software designed to reduce and simplify the work of human coders to increase reliability (*18*).

In each of the two phases, competing teams were given text and metadata for 5,000 publications and single set of metadata on 10,348 data sets of interest, shared between the two phases, for use in training and testing their models. Separate 5,000-publication samples were provided for each phase. The corpus of 10,348 data sets included data maintained by Deutsche Bundesbank and the set of public data sets hosted by the Inter-university Consortium for Political and Social Research (ICPSR). In addition, a single 100-publication development fold was provided separate from the training and testing data to serve as a test for packaging of each team’s model, and as a quick test of their model and the quality of its output[14].

In each phase, an additional separate set of 5,000 publications were held back and used to evaluate the models. After the 1st phase, the phase 1 holdout was also provided to phase 2 competitors to serve as additional training and testing data.

In phase 1, both the train-test publications and the holdout publications were broken into 2,500 publications each that used one or more of the data sets of interest for analysis, as compiled by ICPSR and Bundesbank staff, and 2,500 publications that had not been annotated and had been filtered to not contain data. The data set citations were captured in a separate data set citations JSON file. The citations for the phase 1 train-test publications were provided to competition teams to use as training data, while the citations in the phase 1 holdout were used to test the quality of each team's model in phase 1, and given to teams as additional training data in phase 2.

In phase 2, teams were provided with the phase 1 holdout for additional annotated training data, and then provided with an additional un-annotated set of 5,000 publications to assess their model's behavior on un-curated data. The phase 2 holdout of 5,000 publications was also unannotated, and evaluation of data set detection was based on hand-coded data set reference data revised to make the data more representative of what the models were asked to detect.

Publications

All publication text provided to teams was either open access, and so freely available, or licensed from the publisher for use in the contest by contest participants. In each phase of the competition, a set of publications was provided to the participants and a separate set of publications was held out and kept in reserve so it could be used to evaluate the teams' models. For each publication, participants were provided with PDF and plain text versions of each publication along with basic metadata (pub_date; unique_identifier - DOI or equivalent; text_file_name; pdf_file_name; and publication_id - the unique identifier from our internal system used to manage the data, metadata, and underlying relationships between publications and data sets for the competition).

One particular challenge was that copyright and licensing around research publications limited what publications could be accessed, licensed, and distributed for the competition, and so our universe of publications was limited to publications that were either open access, or published by Sage Publications.

Publication Dataset - Phase 1

- 2500 labeled training publications
- 2500 unlabeled/no-dataset training publications
- 100 publication development fold
- 2500 labeled holdout publications
- 2500 unlabeled/no-dataset holdout publications

In phase 1, 5,000 publications were provided to participants as a train-test data set, 5,000 publications were held back for evaluation, and 100 publications were provided as a separate development fold, for basic model testing and evaluation. The train-test and evaluation holdout each contained 2,500 publications that cited at least one data set, and 2,500 publications that had not been cited by ICPSR as using their data, and had been filtered to not have obvious markers of using data.

The annotated portion of these two sets of publications were drawn from a set of publications provided by Bundesbank that referenced their data and the publications captured in the ICPSR catalog annotated as having used a particular data set for analysis. These publications were collected in a database application designed to facilitate a mix of human and automated content analysis of publications. They were then filtered into two sets: those that were open access, and so could be shared publicly, and those that were not open access, but that were available from our publisher partner (Sage Publications, or “Sage”). Of the 5,100 total publications with annotated data citations provided to phase 1 participants, the 2,550 publications in the train-test corpus (2,500) and development fold (50) were randomly selected from the open access set, so they could be distributed freely to all participants. The 2,500 in the holdout were randomly selected from the remainder of the open access set plus those available from Sage. The un-annotated publications used in phase 1 were all published by Sage - the 2,550 non-annotated publications in the train-test corpus (2,500) and development fold (50) were open access publications from Sage journals. The 2,500 un-annotated publications used in the holdout evaluation corpus were sampled from across Sage Publications’ journal holdings including non-open access journals.

Publication Dataset - Phase 2

- The main publication corpus for phase 2 of the competition was 10,000 unlabeled publications evenly distributed between 6 key topic areas (Education, Health care, Agriculture, Finance, Criminal justice, and Welfare), nicknamed the “wild corpus”.
- 5,000 of these 10,000 were given to teams to work with in phase 2 (randomly selected from within each of the 6 key topic areas to maintain even distribution across topic areas).
- The other 5,000 publications were held out to serve as an evaluation corpus.
- In addition, teams were given the same 100 publication development fold as in phase 1.
- Teams were given the 5,000 publication evaluation corpus from phase 1 to serve as further train-test data.

In phase 2, we worked with Sage to find publications in six key topic areas of interest for partners and future projects (Education, Health care, Agriculture, Finance, Criminal justice, and Welfare). For 28,769 matches, Sage provided PDFs for each and we parsed the text (see details below), removing any that did not parse, or that resulted in file sizes smaller than 20KB, reducing the size of the sample to 25,888. We looked at publication year and type to see if we needed to filter out older publications or non-academic publications, but there were few enough of each class (644 pre–2000 publications and 3,115 non-research articles) that we decided we'd keep all in to preserve as much potential for heterogeneity as possible. From these 25,888 publications, we then randomly selected a total of 10,000 with the goal to keep the distribution across the 6 topic areas equal (so 1666 randomly selected in 2 topic areas, 1667 randomly selected in the other 4). Then, we split the phase 2 corpus to give half to participants and keep half back for evaluation, maintaining equal distribution between the topic areas within each set of 5,000 publications.

Converting PDF files to plain text

The plain text provided for each publication was derived from that publication's PDF file by the competition organizers. It was not intended to be a gold standard, but to serve as an option in case a team preferred not to allocate resources to PDF parsing.

The articles were converted from PDF to text using the open source "pdftotext" application, an Xpdf text extraction system. The basic conversion used the "raw" mode of "pdftotext":

```
pdftotext -raw <path_to_pdf.pdf> <path_to_txt.txt>
```

There are many approaches and tools available for this task. The rationale behind this simplified process for converting pdfs to texts:

1. To render the most usable txt files from available pdfs without over engineering for any specific types of pdf files (e.g., single column vs. multi-column).
2. To have a process that is easily reproducible across different machines for free. That is, not all PDFs convert the same way. Some are more error prone than others. More advanced OCR techniques might have been able to compensate where Xpdf might have fallen short, but relying on more sophisticated and perhaps costly text conversion processes would have made the conversion pipeline more expensive to reproduce and less portable across different applications.

Because of the basic approach, there were some limitations to note:

- Many artifacts from PDF formatting were left behind in the text.
- We had to tweak our processing to get multi-column layouts to output text in order in a linear, single-column text output, and the method we ended up using to achieve this precluded more nuanced processing of other elements of the PDFs.
- Example: tables and charts were not converted in any way to text.

Competition participants were encouraged to try their own conversion process if this text did not meet their needs. If participant teams chose to use another means for converting PDF files to plain text, we asked that they supply us with documentation for installing and running their conversion process so we could start to build up a set of PDF processing strategies that could be reused in the future.

Finding Data Sets

Competitors were provided with two sets of data related to detecting data sets: 1) a catalog of all of the data sets of interest that models were tasked with finding in publications, including basic metadata for all and a list of verbatim mention text snippets for those that were cited in the train-test data; and 2) a subset of these data sets that were actually specifically annotated as having been used for analysis in a given publication.

The data set catalog, provided to participants in the JSON file data_sets.json, contained metadata for all public datasets in the ICPSR data repository and a subset of public data sets available from Deutsche Bundesbank. It includes all data sets sited in the train-test and evaluation corpora, plus many others not cited in either. The data was provided in JSON format for ease of use, a JSON list of JSON objects, each of which contains:

- subjects - list of terms associated with the dataset, based on the [[ICPSR subject thesaurus.]{.underline}](https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/subject)
- additional_keywords - System keyword for where dataset originated.
- citation - Preferred dataset citation.
- data_set_id - Integer ID for dataset from our internal data store of publications, data sets, and relations. This is the identifier used in the data_set_citations.json file to identify relationships between datasets and publications.
- title - Canonical title for dataset.
- name - Canonical title for dataset.
- description - Dataset description, if available.

- unique_identifier - Original unique identifier for dataset, normally a DOI if available.
- methodology - Methodology for dataset, if available.
- date - Date when dataset was published, if available.
- coverages - Geographic coverages, if available.
- family_identifier - Internal system ID, roughly captures datasets that have multiple years but are the same dataset. Inconsistently applied, should not be used in analysis.
- mention_list - Array of strings for annotated mentions as identified by human reviewers. Not an exhaustive list of mentions for any given dataset, and only populated for those data sets cited in the phase 1 train-test corpus.

The mention list is the superset of all unique mention strings associated with each data set across all of that data set's citations where mention data was created. Mention data was only created for data sets cited in the phase 1 train-test corpus.

ICPSR captured when a given data set was used in analysis within a particular publication, but it did not capture particulars on how that determination was made. To provide better data for participants, we implemented a human content analysis protocol to capture mention text for each data set-publication pair included in our train-test corpus (see [Data Set Mention Annotation Process{.underline}](#_23ckvvd) below). Since we manually created this data, given limited time and resources, we initially only did this work for data sets that the teams would be using for training and testing in phase 1. In future work, we intend to provide this kind of information for all data sets of interest, and to refine the protocol to capture the exact position in the text of each mention along with the verbatim text.

Citations of data sets by publications within our phase 1 corpora were captured in separate data_set_citations.json files for each of the train-test and evaluation corpora. Each of these JSON files contains a JSON list of JSON objects, each of which specifics a single relationship between a data set and a publication. This JSON format is also used by models to output detected citations. Each citation contains:

- citation_id - A unique ID for the relationship between one dataset and one publication
- publication_id - Unique ID for a publication which is the same ID for the publication in publications.json
- data_set_id - Unique ID for a dataset which is the same ID for the dataset in the data_sets.json file.
- mention_list - Optional array of strings for alternative references for the dataset in the specific publication (only present in citations included in train-test corpus, and even then, could still

be empty).

- score - Confidence score for the dataset being found in the related publication. In ICPSR-specified citations, the score will be 1.0. In model-created files, will depend on the model.

Even citations from the phase 1 train-test corpus could have an empty mentions list. A given publication could, for example, have been tagged with a dataset by the curator (either at Bundesbank or ICPSR) based on knowledge of the publication and dataset, but a human coder without this knowledge was not subsequently able to find specific mentions within the publication, or the human coder could simply have missed the references. An empty mentions list is not a guarantee that the data set in question was not mentioned.

The list of data sets cited in a particular publication is also not exhaustive. There is the possibility that other data sets from our catalog of data sets of interest were used in analysis within a paper but not captured. The ICPSR data did not include mentions where data was not used in analysis, even of other ICPSR data sets. And named data sets not within our catalog of data sets of interest could also have been used in analysis within a given publication.

Data Set Mention Annotation Process

The ICPSR data contains many explicit ties between publications and data sets that would have been hard to come by otherwise, but the lack of any indication of which parts of the publication indicated the citation relationship made it difficult to identify the linguistic context within the publication that captured the relationship.

To make it easier for participants in the competition to efficiently and systematically engage with the language used to discuss data, we developed a content analysis protocol and accompanying web-based coding application so human coders could examine all of the data set citations in our train-test corpus and capture mention text for each. This required human workers to examine each data set citation in the context of its publication (there were X citations in 2500 training publications) to identify and mark locations in the text where each data set was referenced.

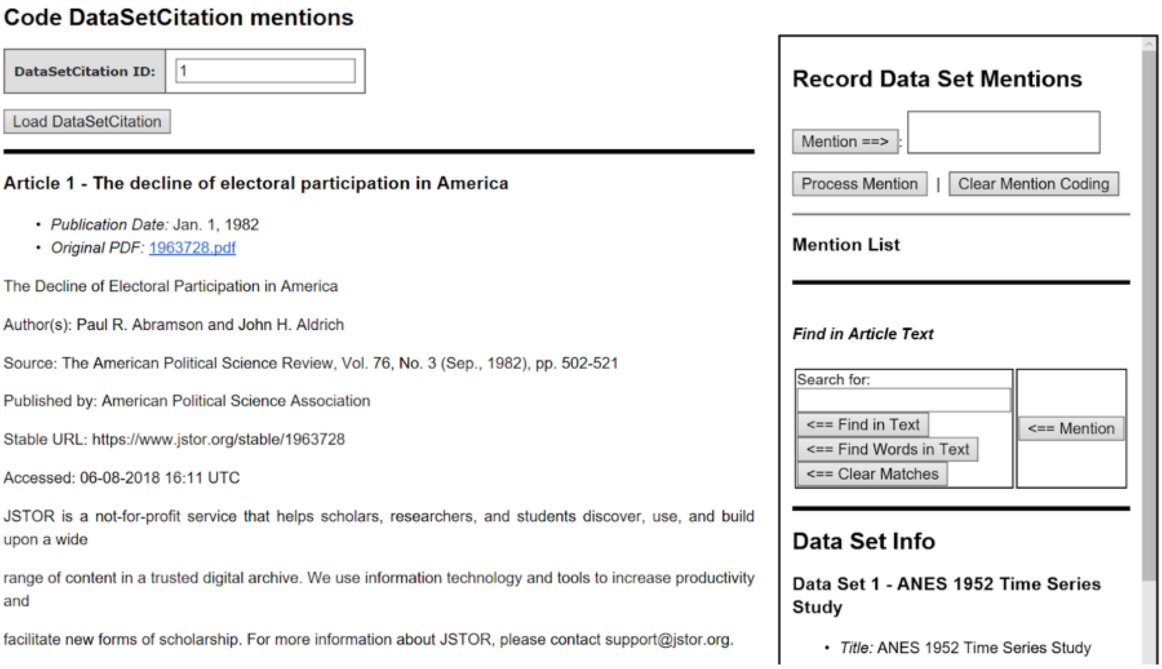
Because of the manual effort required, we only did this for the 2,500 train-test publications that referenced data provided to the teams. We did not manually annotate mention text in the 2,500 publications in the phase 1 holdout, and this made that data a little less useful for teams when it was given to them in phase 2.

Our team of coders was spread across the United States, and so we used a web-based application with a central database store to allow our distributed team of coders to work in parallel. The basic unit of work was a publication-data set pair (so a given publication would be examined as many times as it had different data sets cited within it).

The ICPSR data set repository is very fine-grained in definition of a data set, so each year of an ongoing survey, for example, might have its own data set. To save time, we eventually created the concept of a data set family for these types of data sets and assigned coding for any one instance in a family to all other instances from that family within a given publication. So, for example, multiple years of the same survey or longitudinal data collection were related to each other in a family, and then coding for one year within a paper was used for all other years cited in that paper.

The general process:

- each user was assigned a list of citations to code.
- Once the user logged in to the coding tool, they were presented with a list of the coding tasks assigned to them that included a status of each, so they could track which they had already completed, and a link for each to the coding page.
- Once the user loads a particular citation for coding, they are presented with the following coding page, and are asked to follow the coding instructions in the codebook/documentation for the annotation tool ([[https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit]{.underline}](https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit)):



Article 1 - The decline of electoral participation in America

- *Publication Date:* Jan. 1, 1982
- *Original PDF:* 1963728.pdf

The Decline of Electoral Participation in America

Author(s): Paul R. Abramson and John H. Aldrich

Source: The American Political Science Review, Vol. 76, No. 3 (Sep., 1982), pp. 502-521

Published by: American Political Science Association

Stable URL: https://www.jstor.org/stable/1963728

Accessed: 06-08-2018 16:11 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide

range of content in a trusted digital archive. We use information technology and tools to increase productivity and

facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Data Set Info

Data Set 1 - ANES 1952 Time Series Study

- *Title:* ANES 1952 Time Series Study

{width="6.5in" height="3.486111111111111in"}

Figure 1. The interface of a given publication and a mention capturing process in the coding tool. The left pane contains a full text of an article to code. The right pane contains the coding interface at the top. The "Data Set Info" section contains basic metadata on the data set (title, date of collection, formal identifiers), as well as a list of synonyms gathered so far from publications where the data set is cited.

Coders were instructed to find terms that relate to mentions of the dataset and avoid general synonyms of those terms (for example, tagging "[ANS survey]{.underline}" instead of only "[survey]{.underline}"). If the phrase provides additional information about collection of the dataset, the mention is tagged twice. For example, in the case of "[ANS survey collected/conducted by X]{.underline}", "[ANS survey]{.underline}" is captured first, and then "[ANS survey collected/conducted by X]{.underline}". At the same time, we tried to avoid including too much descriptive information of the dataset - the task is just to code the specific mentions of a particular dataset, including alternate names (e.g. abbreviations, etc.), rather than trying to capture full text in which the data set is discussed.

For more details, including an FAQ that provides guidance on specific issues that arose during coding (like how to deal with data sets that span multiple years), see the content analysis protocol: [[https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit]{.underline}](https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit)

In total, a team of 5 coders, with a background in text analytics for policy research and computational linguistics, completed the task (Emily Wiegand, Neil Miller and Jenna Chapman from Chapin Hall at the University of Chicago, Mengxuan Zhao, Marcos Ynoa and Ekaterina Levitskaya from the CUNY Graduate Center, Computational Linguistics program). The results were then used to re-render data_sets.json and the data_set_citations.json file for the phase 1 train-test data to include mentions.

This combined protocol and tool were developed in-house. Considerations behind building in-house:

- From previous work, we had an open-source tool that did what we would need with minor tweaks, so were able to leverage substantial existing work, though we did have to pay for the work to customize it as well as the AWS t2.large instance on which we hosted it.
- This tool includes templates for human-coding application pages like the one we used, but it is also designed to be used to build up data about publications from multiple sources and this data is straightforward to query and interact with. This allowed us to use the underlying database and application code as the competition dataset database, not just a place to handle mention coding.
- We looked at off-the-shelf text annotators and Qualitative Analysis tool such as lighttag.io, tag.works, NVivo, Atlas.ti, MAXQDA. Unfortunately, given a tight timeline and relatively complex requirements, we didn’t have the time to come up to speed with any of these tools. In addition, we needed the tool to be usable by a distributed team, and that precluded some tools above that did not support distributed workflows.
- For future coding work, we would love to be able to outsource coding tool development, and so are looking at distributed coding applications like lighttag.io and tag.works.

Methods and Fields

For the task of detecting methods and fields for a given publication, our goals were broader than simply providing a vocabulary for each and asking the teams to classify publications against them. We want to encourage development of models that not only can determine when a given publication is a part of an existing field or uses an existing method, but that also understands enough about fields and methods such that they can be used to detect new fields and methods as they emerge, and can then be used to look back through time for traces of these new fields and methods to track their growth and evolution.

To support this goal, we did not give any formal set of either methods or fields that participants needed to train models to classify from. Instead, we provided examples of taxonomies of methods and fields that Sage Publications uses to classify their publications[15], and we directed participants to use them as an example, but to try to make models that would be more creative and potentially able to find new, emerging, or novel fields rather than just fit a publication to a term from a predefined taxonomy.

In practice, this decision to forego any kind of fitting to an existing taxonomy showed the complexity of the problem of understanding fields and methods well enough to detect them based on linguistic context, rather than classifying to an existing vocabulary. Some teams limited themselves to the vocabularies we defined, and the results were uninspiring. Some teams tried to detect based on text, but ended up with a lot of noise and few relevant terms.

In addition, we also learned that there is complexity in “methods” that lumping all methods together did not account for: methods could mean many things, and we started to find sub-categories that we wish we had broken this into: statistical methods, analysis methods, data collection and creation methods, etc.

For future work, for each of these types of information, we intend to first work to decide what exactly we mean by “fields” and “methods”, then find or develop one or more taxonomies to precisely capture what we mean. Once we have these taxonomies, we’ll focus separately on building models to classify publications to them, and making models to extend and update them.

Submission Process

The primary goals of the submission process developed for our competition were:

- to balance the effort needed for a particular group of participants to package their model for submission with the effort needed from the competition organizers to configure, run, and troubleshoot submissions once they were received.
- to begin development of a model packaging strategy that could be used to distribute and allow reuse of any model that uses it.

More specifically, we had the following requirements:

- Create submission infrastructure to make it as straightforward and easy as possible for a team to package their model for submission, including minimizing the understanding needed to use technologies chosen for packaging and deployment and having a built-in way to automatically run the model over the dev fold to validate processing of standard input formats and creation of required output formats.
- Minimize the installation and configuration work needed on part of competition organizers to replicate computing environments as part of model submission process.
- Maximize our ability to see and be able to test how each submission environment is set up, and so avoid accepting a blackbox that could contain anything (including malicious code or sneaky/clever tricks).

Building and Submitting a Model

Our approach for participants building and submitting a model combines Box.com, docker, a git repo for code to implement and support infrastructure, and shell scripts. The central workspace for competition participants was a Box folder that contained example docker files, a copy of the dev fold, and shell scripts that implemented the basic steps of packaging, building, running, and testing a model. The git repository ([[https://github.com/Coleridge-Initiative/rich-context-competition]{.underline}](https://github.com/Coleridge-Initiative/rich-context-competition)) was integral to our framework, but was not used directly by participants. Its code repository was solely used as a home for the code, scripts, and files that made up our submission framework. We did, however, host documentation for participants in the repository's main README and its wiki ([[https://github.com/Coleridge-Initiative/rich-context-competition/wiki]{.underline}](https://github.com/Coleridge-Initiative/rich-context-competition/wiki)).

To get started, participants downloaded a compressed archive of the Box folder and extracted it onto a system with a bash shell. Windows systems were supported, but we recommended that participants with Windows machines work inside a linux virtual machine.

This work folder contained:

- the script “rcc.sh” and its accompanying configuration “config.sh”, that implements all of the basic actions needed to manage docker for a model.
- A set of scaffold files and folders that demonstrate how to hook a model into a docker container, including a Dockerfile with examples of installing OS packages and python packges in a docker container and an example “project” folder with a “code.sh” shell script that

  is called by default when the docker container is run, pre-configured to call a provided example python file named "project.py".
- A copy of the git repo, for use by the scripts.
- A copy of the dev fold, in the standard data folder structure.

The set of scaffold files provided out of the box could be used along with "rcc.sh" to create a simple docker container to test one's local install of docker (including reading from and writing to a data folder configure in "config.sh", running a script in the work folder, and creating output).

Participants were then instructed to work within the "project" folder in their work folder, get their code working first on their local machine, then set up a docker container using the provided example files and get the model running there, to isolate problems with docker from problems with their model.

When participants were ready to submit, they were asked to compress their work folder and upload it to the root of their group's project folder and send an email to the organizers.

Participants were allowed 2 test submissions before the final submission, and most groups took us up on those test submissions in phases 1 and 2. All groups were able to work within the "code.sh" and "project.py" files in "project" to get their model to run, so no further customizations were needed.

Model API

Our submission framework used a file-system based API for giving the model input and accepting output. We interaction through the file system to keep the configuration and implementation simple.

Each time the docker container for a model is run, it is configured to work in a particular data folder.

This data folder has a standard directory structure:

data
| *input*
| | files
| | *text*
| | pdf
|_output

All input information is stored in the “data/input” folder. All output is expected to be stored in the “data/output” folder.The input folder will contain a "publications.json" file, with the same contents as described above in the “Data → Publications” section of this chapter, that lists the articles to be processed in the current run of the model. Publication plain text is stored in “data/input/files/text”, one text file to a publication, with a given publication's text named "<publication_id>.txt". The original PDF files are stored in “data/input/files/pdf”, one PDF file to a publication, with a given publication's text named "<publication_id>.pdf".

The output folder starts out empty, and is where the model is expected to place 4 output files after each run of the model:

- **data_set_citations.json** - A JSON file that contains publication-dataset pairs for each detected mention of any of the data sets provided in the contest data_sets.json file. The JSON file should contain a JSON list of objects, where each object represents a single publication-dataset pair.
- **data_set_mentions.json** - A JSON file that should contain a list of JSON objects, where each object contains a single publication-mention pair for every data set mention detected within each publication, regardless of whether a gvien data set is one of the data sets provided in the contest data set file.
- **methods.json** - A JSON file that should contain a list of JSON objects, where each object captures publication-method pairs.
- **research_fields.json** - A JSON file that should contain a list of JSON objects, where each object captures publication-research field pairs.

Running a Submitted Model

Once a model was submitted, the competition organizers followed a standard script for running the model and processing its output for analysis:

- For each submission, an AWS instance was spun up from a standard image pre-configured to run models built using our submission framework.
- The evaluator connected to the instance and started a screen session, so work would not be disrupted if connection to server was lost.
- The model was downloaded to the server and extracted.

- The submission container was built on the server using the provided Dockerfile and “rcc.sh”, and then the container was run over the dev fold to test basic functionality of the container and the model, and to give an estimate of time needed to complete.
- Once the dev fold was successfully processed, “config.sh” was reconfigured to point at the evaluation corpus, and the model was run over the evaluation corpus.
- Once the model completed, standard evaluation Jupyter notebooks in the git submission framework repository were configured to the current projects output and run to generate materials for judges to evaluate the submission.
- Output and results were copied to a central storage area, and the instance used to run the model was terminated.

Throughout this process, the evaluator communicated any problems with the participant team and worked with the team to address problems and turn around a new version of the model as quickly as possible. If a team’s model performed poorly on the standard size machine, we also would sometimes try different sizes of server to give them an idea of whether their problem was related to needing more compute power, or was a limitation of their approach independent of available resources.

Notes on the Submission Process

We chose Box.com because we have unlimited space there through NYU, and so we were able to accommodate not only whatever data participants needed to provide to make their models work, but also all of the data we provided to participants for training and testing. To minimize confusion, we pre-configured and shared each team’s Box folder with them, so they did not have to do any setup.

To setup the infrastructure in each folder, we created a git repository (https://github.com/Coleridge-Initiative/rich-context-competition) that contained all of the files, shell scripts, and templates needed to: 1) configure a new instance of a team folder, for use by competition staff setting up team folders; 2) develop, package and test deployment of a model (participants); and 3) support building, running, and evaluating the models once they were submitted.

We considered using github to store participant submissions, but chose Box because of its unlimited storage.

We considered using an external service like CodaLab or Kaggle, but an initial assessment of each suggested that they would not meet our needs without substantial changes to the design of our competition:

- Codalab looked promising, but its documentation was sparse and our time frame was short enough that we weren't comfortable we could get up to speed with it quickly enough to make a reliable, easy-to-use competition with it.
- Kaggle seemed designed for more basic competition designs (our evaluation steps were fuzzy, so couldn't just take their outputs and make scores - this is not entirely a knock on them - it would be great to get our tasks to the point where they fit in this framework, we just don't have the data yet), and there were also licensing complications we weren't comfortable sorting out. We also needed control over manual evaluation and were concerned there that their submission and evaluation system wouldn't support the bespoke nature of our submissions.
- For both, we also simply weren't comfortable that we'd be able to get up to speed on the platform in time to make the experience of participating in the competition as pleasant and painless as possible.

We also wanted to have the flexibility to run many models in parallel and give models substantial resources if needed, to see how they performed with different magnitudes of computing resources and to allow us to try to throw raw compute power at a model if it was running too slowly, to get it to complete so we could give as good of feedback as possible. We not only wanted groups to be able to do preliminary submissions, but we wanted to make sure we could give as much feedback as possible. This led us toward a container-based approach where we did what we could to abstract and simplify the running of models, and allowed for flexibility and configurability in the instances that we spun up to run the models.

Evaluation

In both phases of the competition, we evaluated raw mentions, research fields, and research methods separate from citation of named data sets.

Phase 1 Evaluation

Mentions, Methods and Fields

In phase 1, expert social science judges evaluated mentions, methods, and fields in two ways: 1) we randomly selected 10 publications to manually examine each team's output against, and made notes of good and bad for each team, then ranked the teams within each publication; and 2) we generated distributions of all values found across all publications within each type of value, counted the occurrences of each, compared the distributions across teams, and ranked the teams based on how their distributions compared. To create overall rankings, the judges met, compared notes and individual rankings, and then agreed on an overall ranking of the teams.

Data Set Citations

To evaluate data set citations in phase 1, we used the ICPSR citation data as our evaluation baseline for creating a confusion matrix based on how each team's citation findings compared to ICPSR's baseline, and we calculated precision, recall, and F1 scores from the confusion matrix to compare across teams. To create the confusion matrix for each team, we started with a list of all of the data set-publication pairs found either in ICPSR's baseline or the team's output. We created found-or-not (1 or 0) vectors for every publication-data set pair for the baseline, and for the team. Then, for each data set-publication pair, we compared the values between the baseline vector and the team vector to decide how to update the confusion matrix for that pair: if a team agreed with ICPSR on presence of a data set, that was counted as a true positive (TP). If the team found a data set that ICPSR did not, that was counted as a false positive (FP). If a team missed a data set ICPSR indicated was present, it was counted as a false negative (FN). We did not develop a way to capture true negatives since the metrics we used to evaluate did not require it. In addition, as part of the processing to create the overall confusion matrix, we created per-publication confusion matrices for each publication, so we could track average false positives and false negatives per publication, and highlight publications that were higher than the average, for more detailed evaluation.

We also deferred figuring out “mentioned” vs. “used in analysis” in our initial competition, to make the initial task more manageable. This decision, combined with the traits of the ICPSR data, caused substantial noise in the phase 1 precision/recall/F1 scores. For example, even models that figured out that a longitudinal data set was present sometimes got many false positives and false negatives because they got the years wrong, and models that correctly found ICPSR data sets used in discussion had those counted as false positives because ICPSR had only captured data sets used in analysis.

Phase 2 Evaluation

In evaluating phase 2, we kept the division between mentions, fields, and methods and citations, but we refined our evaluation methods in based on what we'd learned in the first phase.

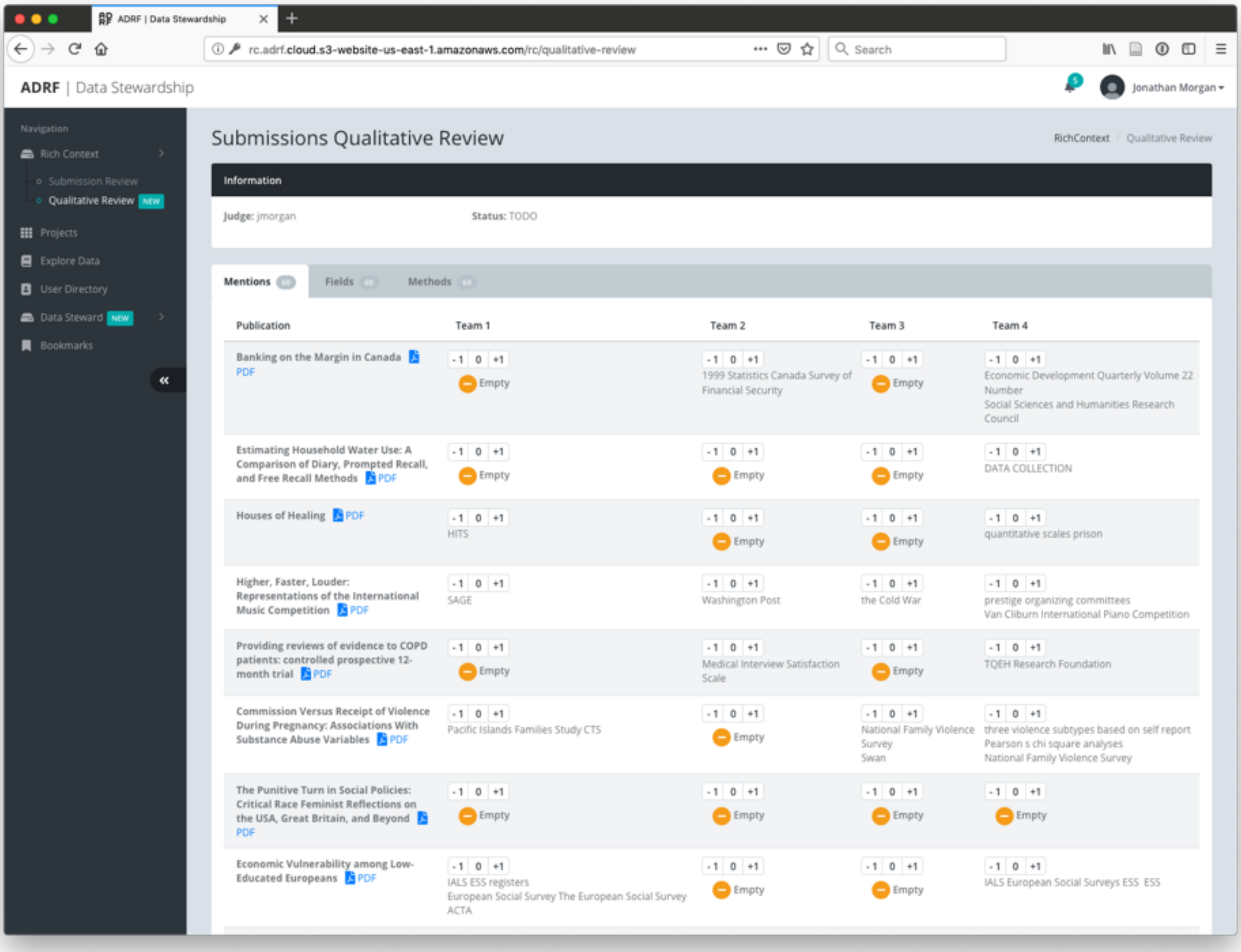
Mentions, Methods and Fields

For mentions, methods, and fields in phase 2, we kept the basic strategy of: 1) comparing the values created by each team's model in the context of a set of selected publications and 2) reviewing the overall distributions of values for each team.

We expanded the number of publications across which we compared values to make the sample reviewed more representative, though, and created a web-based tool to help judges deal with the added work from more publications to review. We also selected publications differently for data mentions from fields and methods, choosing publications with different levels of agreement between the teams on whether data was present or not, to start to evaluate the different model's ability to detect data at all, in addition to comparing the results when they thought a publication contained data.

For fields and methods (and data set citations), we selected 20 publications for each of our 6 topic areas of interest (Education, Health care, Agriculture, Finance, Criminal justice, and Welfare) with a few extras (2 extra in finance and 1 extra in criminal justice), for a total of 123 publications to compare values across. Within the 20 publications per topic area, we worked through a random selection of articles picking publications to add to our sample to fill out a rough ratio within each topic area of 5:4:1 between publications with titled data sets (5); data described, but not titled (4); and no data (1).

To make it easier for the judges to work through this increased number of publications, we also created a tool that collected the output for each team side-by-side per publication along with a link to each publication's PDF, and had a place for the judge to score each team's output for a given publication from among "–1", "0", and "1". Once judges scored all output, we then created rankings based on the sum of each team's scores.



Submissions Qualitative Review

Information

Judge: jmorgan Status: TODO

Publication	Team 1	Team 2	Team 3	Team 4
Banking on the Margin in Canada PDF	-1 0 +1 Empty	-1 0 +1 1999 Statistics Canada Survey of Financial Security	-1 0 +1 Empty	-1 0 +1 Economic Development Quarterly Volume 22 Number Social Sciences and Humanities Research Council
Estimating Household Water Use: A Comparison of Diary, Prompted Recall, and Free Recall Methods PDF	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 DATA COLLECTION
Houses of Healing PDF	-1 0 +1 HITS	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 quantitative scales prison
Higher, Faster, Louder: Representations of the International Music Competition PDF	-1 0 +1 SAGE	-1 0 +1 Washington Post	-1 0 +1 the Cold War	-1 0 +1 prestige organizing committees Van Cliburn International Piano Competition
Providing reviews of evidence to COPD patients: controlled prospective 12-month trial PDF	-1 0 +1 Empty	-1 0 +1 Medical Interview Satisfaction Scale	-1 0 +1 Empty	-1 0 +1 TQEH Research Foundation
Commission Versus Receipt of Violence During Pregnancy: Associations With Substance Abuse Variables PDF	-1 0 +1 Pacific Islands Families Study CTS	-1 0 +1 Empty	-1 0 +1 National Family Violence Survey Swan	-1 0 +1 three violence subtypes based on self report Pearson s chi square analyses National Family Violence Survey
The Punitive Turn in Social Policies: Critical Race Feminist Reflections on the USA, Great Britain, and Beyond PDF	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 Empty
Economic Vulnerability among Low-Educated Europeans PDF	-1 0 +1 IALS ESS registers European Social Survey The European Social Survey ACTA	-1 0 +1 Empty	-1 0 +1 Empty	-1 0 +1 IALS European Social Surveys ESS ESS

{width=“6.5in” height=“5.013888888888889in”}

Figure 2: The interface given to judges to evaluate data set mentions, research fields, and research methods.

For manual evaluation of data set mentions, we used the same tool described above, but we chose a different sample of 60 publications based on agreement between the output of the different participant team models as to whether publications had data mentions. To generate this sample, we first loaded all of the output from each team’s model into our work database. We then made a list of all of the publications in our phase 2 holdout and, for each publication, the count of teams that had data set mentions for that publication. We then sampled to get 60 publications:

- 10 publications where all teams agreed there was no data.
- 10 publications where all teams agreed there was data.
- 40 publications where the teams disagreed on whether there was data.

For the 40 publications with disagreement, we selected publications with 1 team, 2 teams, and 3 teams agreeing data was present proportional to the distribution of each level of agreement in the broader sample:

- 17 from 1 (1439/5000 = 0.2878; 0.2878 * 60 = 17.268)
- 20 from 2 (1741/5000 = 0.3482; 0.3482 * 60 = 20.892)
- 13 from 3 (1080/5000 = 0.216; 0.216 * 60 = 12.96)

We then asked a separate pair of qualitative judges to use the tool to compare and evaluate the data set mentions generated by the teams across these publications.

Data Set Citations

Our analysis of data set citations in phase 2 required a more substantial rethinking since we did not have any starting point for presence or absence of data like the ICPSR corpus. We implemented a method of creating a confusion matrix that could be used to generate precison, recall, and F1 scores more closely aligned with the task we'd assigned the teams to implement - finding mentions of data and data sets within publications.

To implement this, we started with the sample of 123 publications used for evaluating mentions and fields above and:

- Captured all "data references" within each of those publications using a new human coding protocol. This included external titled data sets either discussed or used in analysis, external data without a title that was discussed or used in analysis, and data created by the researcher for a given study.
- For each data reference, we compared all mentions and citations created by each team for the publication to the information on the data reference within that publication and marked any that were "related" to the data reference.
- Finally, we used the list of references as a baseline and built a confusion matrix based on whether each team had found mentions or citations "related" to each of the data references, along with a "false positive" record where the baseline was always 0 and the team was assigned a 1 if they had one or more mentions or citations that were not "related" to any data reference.

Capturing Data References

To capture data references in our sample of publications, we created a basic protocol for an initial round of data creation ([[https://docs.google.com/document/d/1aFPEtT4hd93kcsOEzocyB6-a4Hu8WcemKTld–98Q25k/edit#heading=h.f3u3kdbg87s4]{.underline}](https://docs.google.com/document/d/1aFPEtT4hd93kcsOEzocyB6-a4Hu8WcemKTld–98Q25k/edit#heading=h.f3u3kdbg87s4)), then evaluated the results throughout the rest of the process. We used a single data reference coder to encourage consistency in output. Our data reference coder worked within a spreadsheet to, for each publication in our sample:

- Flag all paragraphs where data was mentioned.
- Cluster mentions together that refer to a single dataset.
- Give each cluster of mentions a row in the spreadsheet. These are our “data references”.
- Then, for each data reference:
    - Collect all mentions that refer to the reference.
    - decide if the data set is simply cited (“cited”), or if it is one used in analysis (“analysis”) in the publication
    - Capture words or phrases that are key to identification as “key terms”.
    - Also capture any broader contextual text in “Context”, so it could be used to better understand the nature of the “data reference”.
    - If data set title is present, capture it.
    - Try looking up the data set in the database, and if it is there, store its data set ID.

We tried to capture detailed context on each reference for a couple reasons: 1) To make it easier for reviewers of this data to evaluate the quality of each data reference; 2) To give more context for judges deciding if mentions and citations for a given team were “related” to a given data reference.

Finding Related Mentions and Citations

After the data references were captured, a team of coders then looked at each data reference related to the selected publications for each team to see if data set citations and mentions by the team were “related” to the data reference.

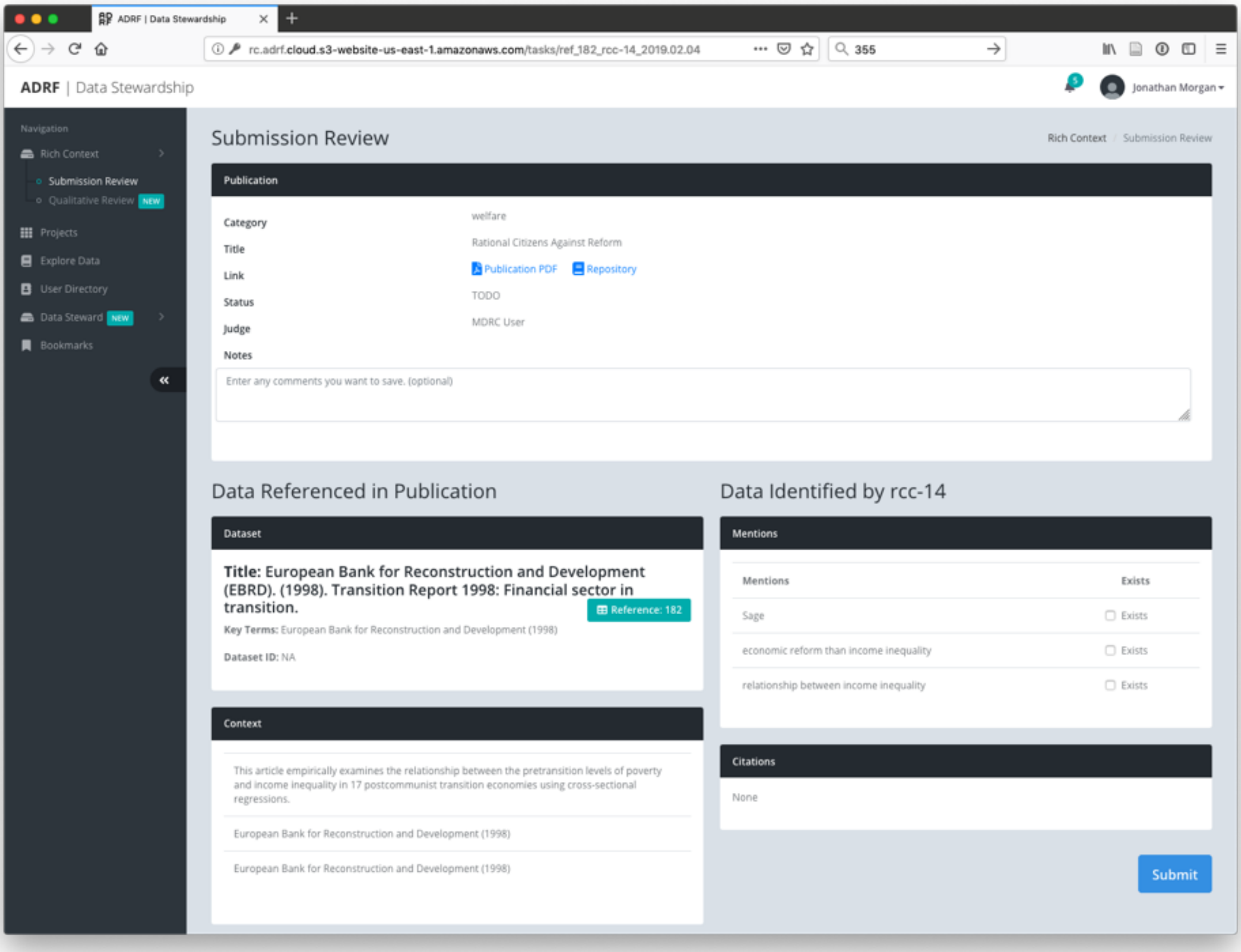
The coders, subject matter experts in the different key topic areas, looked at each “data reference” in publications in their area of expertise. For each, they evaluated it against the mentions and citations output by the model of each team that found mentions or citations in the selected publication. For each reference-team pair, the coder flagged any mentions or citations they deemed “related to” the current data reference.

In our protocol
([[https://docs.google.com/document/d/1Hi13N6gfiRz9nfwCoUQrey8v_ozY7fKHMtHV4GgX2ys/edit#]{.underline}]
(https://docs.google.com/document/d/1Hi13N6gfiRz9nfwCoUQrey8v_ozY7fKHMtHV4GgX2ys/edit))
,
we describe the coding task as “When you are judging data mentions, we want to mark mentions on the right as ”exists“ if they are related to the data referenced on the left, and make sure to not mark any mentions as "exists" that are not related.”, balanced with “If in doubt, don’t mark a given mention as related.”

The definition of “related to” is purposely fuzzy. Our goal was to give credit for finding language related to a dataset even if it wasn’t a perfect, formal reference, but to also make sure to not mark things that are obviously unrelated. To help to flesh this distinction out, we gave examples and analogies and training, and we had coders work through a few data references on their own then discuss their decisions.

An example from the protocol: “Think of it as a fuzzy match - we want to give the models the benefit of the doubt if they get close, especially if they detect some but not all key terms or phrases or find a mention of the basic type of data a named data set represents (”wage data“ for IDES Unemployment Wage Records, for example), but we also want to make sure to reject things that are obviously not related.”

Coders used a web-based coding tool that listed out their assigned coding tasks and pulled together all of the information so they just had to scan the page, open the associated PDF if they had questions, and then mark related items and Submit to save their coding:



Submission Review

Publication

Category	welfare
Title	Rational Citizens Against Reform
Link	Publication PDF Repository
Status	TODO
Judge	MDRC User
Notes	

Enter any comments you want to save. (optional)

Data Referenced in Publication

Dataset

Title: European Bank for Reconstruction and Development (EBRD). (1998). Transition Report 1998: Financial sector in transition. Reference: 182

Key Terms: European Bank for Reconstruction and Development (1998)

Dataset ID: NA

Context

This article empirically examines the relationship between the pretransition levels of poverty and income inequality in 17 postcommunist transition economies using cross-sectional regressions.

European Bank for Reconstruction and Development (1998)

European Bank for Reconstruction and Development (1998)

Data Identified by rcc-14

Mentions

Mentions	Exists
Sage	□ Exists
economic reform than income inequality	□ Exists
relationship between income inequality	□ Exists

Citations

None

{width="6.5in" height="5.0138888888888889in"}

Figure 3: The interface given to judges to evaluate whether a given team's data set mentions and citations were related to a given data reference.

As one would expect, while we got coders on the same page, each had subtly different ideas about what was or was not "related to". To remove some of this variability from our final data, we then had a sole experienced researcher who understood what we were trying to do review all coding and, when he saw coding that obviously did not fit his understanding, either: revise to fit his understanding of "related to"; or flag as one he was unsure of and note his thoughts.

This experienced researcher also served as a final reviewer of the data references that were collected, marking any that did not actually refer to data as needing to be removed from our final analysis.

Finally, the protocol designer reviewed all removed data references, corrections, and ambiguities flagged for additional review, and made a final set of corrections.

Scoring the Results

To create a “related to” confusion matrix for each team, we started with a list of all of the data references that our final reviewers indicated should be included in our analysis (165 total). We created found-or-not (1 or 0) vectors with a value for every reference set to 1 for the baseline, and then set based on our coding for each team. For each publication, we also included a false positive item that was always 0 for the baseline, and that was set to 1 for a given team if they had any mentions or citations that were not “related to” a data reference from that publication.

To build a given team’s vector, for each data references, we checked to see if any of the team’s mentions or citations had been marked as “related to” that reference. If one or more of the team’s mentions or citations was marked as “related to”, we gave that reference a “1” for that team. If not, we gave it a “0”. Then, for each publication’s false positive item, if the team had 1 or more mentions and/or citations that were not “related to” any data reference, the team got a “1” for that entry. If not, they got a “0”.

To build out a confusion matrix, we went reference by reference: If the team found mentions and/or citations related to the reference, that was counted as a true positive (TP). If a team did not have any mentions or citations related to a given data reference, it was counted as a false negative (FN). Then, for the publication, if the team had 1 or more mentions and/or citations that were not “related to” any data reference, this was counted as a false positive (FP).

We did not develop a way to capture true negatives since the metrics we used to evaluate did not require it.

Discussion

Given the time and resources available to put the competition together, the competition’s design was effective, but required some iteration within each of the phases. We modified and updated both training data and model submission infrastructure in response to participant feedback, and the participants were generally quite positive about the experience.

The docker-based model submission process worked well for the competition, but subsequent use of the models by Digital Science and Bundesbank has revealed a need to more precisely design how the models work within their docker container and the APIs they provide so packaged models implement a more re-usable API. For example, to be readily able to be used within an existing environment, the model needs to be able to be invoked from a simple unit of code (a python function, for example), rather than needing to spin up an instance of a container each time you want results.

To facilitate re-use, we need much more detailed specification of how the participants should implement their models. For example:

- If a submission is implementing multiple tasks, each should be broken into its own separate API so it can be used separately (so separate services for mention detection, field detection, and data detection).
- We need to better specify how we expect the models to be re-trained, in particular elements of the model we expect to be easily changed and which we expect would require a full retraining to tune. For example, we hoped to be able to easily switch out the data sets of interest that are detected specifically without needing to retrain on a full corpus referring to those data sets, but we didn’t mention this, and none of the models worked this way.

In terms of community building, we inspired participation and the workshop and discussions after the competition lead to collaborations between pairs of sponsors and participants and collective work on making a gold standard corpus that could be used to develop better models in the future (a great step toward higher quality models), but we need to continue to work to nurture and grow this community.

The data for the competition was a great start, but trying to use it to detect data mentions and then start to get at whether data was simply discussed or actually used in analysis revealed how much work remains to make high quality training data. The base ICPSR data did not include mention text where we did not create it, and so for the majority of data sets, the only text available for characterizing a data set was the title and a paragraph of description, no examples of how the data would be discussed within a publication. It also did not capture non-ICPSR data sets, nor did it include data sets mentioned but not used in analysis. We need to be able to work with imperfect data, but the complexity of this task makes it a good fit for better training data. We also found that our definition of a data set was too specific – ICPSR is granular down to the year of some of their formal data collections. Data signatures of interest in the real world might just be clusters of key terms without a formal title, and our data and models need to account for this.

Our evaluation approaches were effective given the time we had, but they also had significant limitations. In phase 1, the ICPSR data was great for a model that finds named data sets used in analysis, but it was not as good a fit for evaluating models trying to detect data citations in general. For example, some high quality models were scored with many false positives that, on review, were actually correct, but for non-ICPSR data sets.

In phase 2, our design and evaluation data creation attempted to account for the limitations of phase 1 - to move from just looking at titled ICPSR and Bundesbank data sets used for analysis and begin to look at all the ways data is discussed in academic papers, and how much of that discussion the combined mentions and citations of each team was aware of. Its effectiveness depended on how well we designed and carried out each of these three steps.

We are comfortable with the quality of the resulting data, but it should be noted that given the time and resources available to us, we had to make a choice between quality of data and reproducibility. In a perfect world, content analysis is the discipline of reliably being able to use a well-designed protocol to create content of comparable quality regardless of who does the coding. Given this project's relatively tight timelines and limited resources, in this process we prioritized quality of data over reproducibility. We created relatively detailed coding protocols for each step of the process and we designed review and refinement into our processes, but we did not have time to go through multiple rounds of training and evaluation to make each of the protocols reliable and reusable. At the end, we introduced consistency by having experienced researchers familiar with our goals review all the output and either correct problems or flag items that should not be used in analysis. We believe that this created a reasonable level of consistency and quality in our output, but intend to refine these protocols for use in the future,

Conclusion

Given the time and resources available, we consider the competition design to have been effective. The design attracted letters of intent from 20 teams from 8 countries and 12 teams actually submitted code. The models were interesting and some of the solutions were novel and surprisingly effective given their novelty. A nascent community of practice was also formed. Discussions after the competition led to collaborations between participants and collective work on making a gold standard corpus that could be used to develop better models in the future – an important step toward higher quality models. The models also ended up being re-usable as they are, though in a limited scope, and at least one sponsor has been able to run them and get useful output.

Additional work is continuing in three directions. The first is developing better corpora: that is discussed in the Appendix to this chapter by Alex Wade and Sebastian Kohlmeier. The second is developing a community of practice: a recent workshop (https://coleridgeinitiative.org/richcontextworkshop) took the next step to doing so. The third is to further develop the machine learning and natural language processing tools through broader based competition: this is discussed in more detail in the concluding chapter in this book.

Appendix - Standardized Metadata, Full Text and Training/Evaluation for Extraction Models

Key challenges when working on an NLP task like dataset mention extraction that requires access to scholarly literature include the proliferation of metadata sources and sourcing of full text content. For example, each metadata source has their own approach for disambiguation (e.g. recognizing that A. Smith and Anna Smith are the same author) or de-duplication of content (clustering pre-prints and final versions into a single record). As a result competition organizers and NLP researchers currently use ad-hoc processes to identify metadata and full text sources for their specific tasks which results in inconsistencies and a lack of versioning of input data across competitions and projects.

One way these challenges can be addressed is by using a trustworthy metadata source like [[Semantic Scholar's open corpus]{.underline}](http://api.semanticscholar.org/corpus/) developed by the Allen Institute for Artificial Intelligence (AI2) or [[Microsoft's Academic Graph]{.underline}](https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema)
that make it easy to access standardized metadata from an openly accessible source. In addition, both Semantic Scholar and the Microsoft Academic Graph provide topics associated with papers which makes it easy to narrow down papers by domain. If full text is needed we recommend tying the metadata to a source of open access full text content like [[Unpaywall]{.underline}](https://unpaywall.org/data-format) to ensure that the full text can be freely redistributed and leveraged for model development.

To gather the data we recommend collecting a sufficiently large set of full text papers (3,000–5,000 minimum) with their associated metadata and providing participants with a standardized format of the full text. More data might be required if data is split across many scientific domains. For example for a task like dataset extraction, reference formatting is often inconsistent across domains and dataset mentions can potentially be found in different sections (e.g. background, methods, discussion, conclusion or the reference list) throughout the text. Once a decision has been made on the full text to include, the PDF content can be easily converted into text in a standardized format using a PDF to text parser like [[AI2's ScienceParse]{.underline}](https://github.com/allenai/spv2) (which handles key tasks like metadata, section heading and references extraction).

Once the metadata and full text dataset has been created it can be easily versioned and used again in future competitions. For example, if updated metadata is needed it's easy to go back to the original metadata source (for example by using Semantic Scholar's [[API]{.underline}](http://api.semanticscholar.org/)) to get the latest metadata.

[Annotation Protocols to Produce Training & Evaluation Data]{.underline}

A common approach to machine learning known as **supervised learning** uses labelled, or annotated, data to train a model what to look for. If labelled data is not readily available, human annotators are frequently used to label, or code, a corpus of representative document samples as input into such a model. Different labelling tasks may require different levels of subject domain knowledge or expertise. For example, coding a document for different parts of speech (POS) will require a different level of knowledge than coding a document for mentions of upregulation of genes. The simpler the labelling task, the easier it will be for the coders to complete the task, and the more likely the annotations will be consistent across multiple coders. For example, a task to identify a *mention of a dataset* in a document might be far easier than the task of identifying only the*mentions of datasets that were used in the analysis phase of research*.

In order to scale the work of labelling, it is usually desirable to distribute the work amongst many people. Generic crowdsourcing platforms such as Amazon's Mechanical Turk can be used in some labelling exercises, as can more tailored services from companies such as TagWorks and Figure-Eight. Whether the labelling is done by one person or thousands, the consistency and quality of the annotations needs to be considered. We would like to build up a sufficiently large collection of these annotations and we want to ensure that they are of a high quality. How much data needs to be annotated depends on the task, but in general, the more labelled data that can be generated the more robust the model will be.

As mentioned above, we recommend 3000–5000 papers, but this begs the question of how diverse the subject domains are within this corpus. If the papers are all within from the finance sector, then a resulting model might do well in identifying datasets in finance, but less well in the biomedical domain since the model was not trained on biomedical papers. Conversely, if our 3000–5000 papers are evenly distributed across all domains, our model might be more generically applicable, but might do less well over all since it did not contain enough individual domain-specific examples.
As a result, we recommend labelling 3000–5000 papers within a domain, but we plan to do so in a consistent manner across domains so that the annotations can be aggregated together. In this manner, as papers in new domains are annotated, our models can be re-trained to expand into new domains. In order to achieve this, we intend to publish an open annotation protocol and output format that can be used by the community to create additional labelled datasets.

Another factor in deciding the quantity is the fact that the annotations will be used for two discrete purposes. The first is to *train* a machine learning model. This data will inform the model what dataset mentions look like, from which it will extract a set of features that the model will use and attempt to replicate. The second use of the annotations is to *evaluate* the model. How well a model performs against some content that it has never seen before. In order to achieve this, labelled data are typically split randomly into training and evaluation subsets.

One way to evaluate how well your model performs is to measure the **recall** and **precision** of the model's output, and in order to do this we can compare the output to the labelled evaluation subset. In other words, how well does our model perform against the human annotations that it was not trained on and has never seen. Recall is the percentage of right answers the model returned. For example, if the evaluation dataset contained 1000 mentions of a dataset, and the trained model returned 800 of them, then the recall value would be .80. But what if the model returned everything as a dataset, then it would get all 1000, plus a whole bunch of wrong answers. Obviously, the precision of the model is important too. Precision is the percentage of answers returned that were right. So, continuing the example above, if the model returned 888 answers, and 800 of those were right, then the precision of the model would be ~.90. But again, if the model returned only one right answer and no wrong ones, the precision would be perfect. So, it is important to measure both precision and recall. In summary, the model in this example, got 80% of the right answers, and 90% of the answers it returned were right. The two measures of recall and precision can be combined into an F1 score of ~.847.

If we then make modifications to our model, we can re-run it against the evaluation dataset and see how our F1 score changes. If the score goes up, then our new model performed better against this evaluation data. If we want to compare several different models to see which one performed best, we can calculate an F1 score for each of them. The one with the highest F1 score has performed the best. Consequently, the quality of the annotations are critical for two reasons: first, the accuracy of a *model* will only be as good as the data upon which it was trained. And secondly, the accuracy of the *evaluation* (in this case the F1 score) can be affected by the quality of the data it is evaluated against.

References

1. D. A. Murdick, Foresight and understanding from scientific exposition (FUSE). *Retrieved June*. **21**, 2015 (2011).2. D. Murdick, “Finding Patterns of Emergence in Science and Technology” (OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE WASHINGTON DC INTELLIGENCE ADVANCED RESEARCH PROJECTS ACTIVITY, 2012).3. J. Perrie *et al.*, Implementing Recommendation Algorithms in a Large-Scale Biomedical Science Knowledge Base. *arXiv Prepr. arXiv1710.08579* (2017).4. D. Brickley, M. Burgess, N. Noy, in *The World Wide Web Conference* (ACM, 2019), pp. 1365–1375.5. S. Ovadia, ResearchGate and Academia. edu: Academic social networks. *Behav. Soc. Sci. Librar.* **33**, 165–169 (2014).6. I. Soboroff, I. Ounis, C. Macdonald, J. J. Lin, in *TREC* (Citeseer, 2012), vol. 2012, p. 20.7. M. Kremer, H. Williams, Incentivizing innovation: Adding to the tool kit. *Innov. policy Econ.* **10**, 1–17 (2010).8. B. D. Wright, The economics of invention incentives: Patents, prizes, and research contracts. *Am. Econ. Rev.* **73**, 691–707 (1983).9. H. L. Williams, Innovation Inducement Prizes: Connecting Research to Policy. *J. Policy Anal. Manag.* **31**, 752–776 (2012).10. T. Kalil, *Prizes for technological innovation* (Brookings Institution Washington, DC, 2006).11. K. Boudreau, K. Lakhani, How to manage outside innovation. *MIT Sloan Manag. Rev.* **50**, 69 (2009).12. J. Mateos-Garcia, W. E. Steinmueller, The institutions of open source software: Examining the Debian community. *Inf. Econ. Policy*. **20**, 333–344 (2008).13. B. M. Sadowski, G. Sadowski-Rasters, G. Duysters, Transition of governance in a mature open software source community: Evidence from the Debian case. *Inf. Econ. Policy*. **20**, 323–332 (2008).14. W. W. Chapman *et al.*, Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions (2011).15. A. Tong *et al.*, Overview of the NIST 2016 LoReHLT evaluation. *Mach. Transl.* **32**, 11–30 (2018).16. L. Wissler, M. Almashraee, D. M. Díaz, A. Paschke, in *IEEE GSC* (2014).17. D. Riff, S. Lacy, F. Fico, B. Watson, *Analyzing media messages: Using quantitative content analysis in research* (Routledge, 2019).18. S. C. Lewis, R. Zamith, A. Hermida, Content analysis in an era of big data: A hybrid approach to computational and manual methods. *J. Broadcast. Electron. Media*. **57**, 34–52 (2013).

Chapter 6 - Finding datasets in publications: The Allen Institute for Artificial Intelligence approach

author:
- |
Daniel King, Waleed Ammar, Iz Beltagy, Christine Betts
Suchin Gururangan and Madeleine van Zuylen

`Allen Institute for Artificial Intelligence, Seattle, WA, USA`
`daniel  @allenai.org`

bibliography:
- 'acl2015.bib'

title: Finding datasets in publications: The Allen Institute for AI Approach

Introduction

The Allen Institute for Artificial Intelligence (AI2) is a non-profit research institute founded by Paul G. Allen with the goal of advancing artificial intelligence research for the common good. One of the major undertakings at AI2 is to develop an equitable, unbiased software platform Semantic Scholar[1] for finding relevant information in the scientific literature. Semantic Scholar extracts meaningful structures in a paper (e.g., images, entities, relationships) and links them to other artifacts when possible (e.g., knowledge bases, GitHub repositories), hence our interest in the rich context competition (RCC). In particular, we participated in the RCC in order to explore methods for extracting and linking datasets used in papers. At the time of this writing, Semantic Scholar comprehensively covers the computer science and biomedical literature, and we plan to expand our coverage in 2019 to other scientific areas, including social sciences.

In the following sections, we describe our approach to the three tasks of the RCC competition, which are described in more detail in Chapter 5:
1. extracting the datasets used in publications,
2. predicting the field of research of publications
3. extracting the methods used in publications

Methods

Dataset Extraction and Linking

This task focuses on identifying datasets used in a scientific paper. Datasets which are merely mentioned but not used in the research paper are not of interest. This task has two sub-tasks:

1. Citation prediction: extraction and linking to a provided knowledge base of *known datasets*, and
2. Mention prediction: extraction of both *known and unknown* dataset mentions.

Provided Data

The provided knowledge base of known datasets includes approximately 10K datasets used in social science research. The high textual similarity between different datasets in the knowledge base informs our approach for linking dataset mentions to their dataset in the knowledge base. Approximately 10% of the datasets in the knowledge base were linked one or more times in the provided corpus of 5K papers. To attempt to generalize mention discovery beyond those present in the knowledge base, we train a named entity recognition (NER) model on the noisy annotations provided by the labeled mentions in the knowledge base.

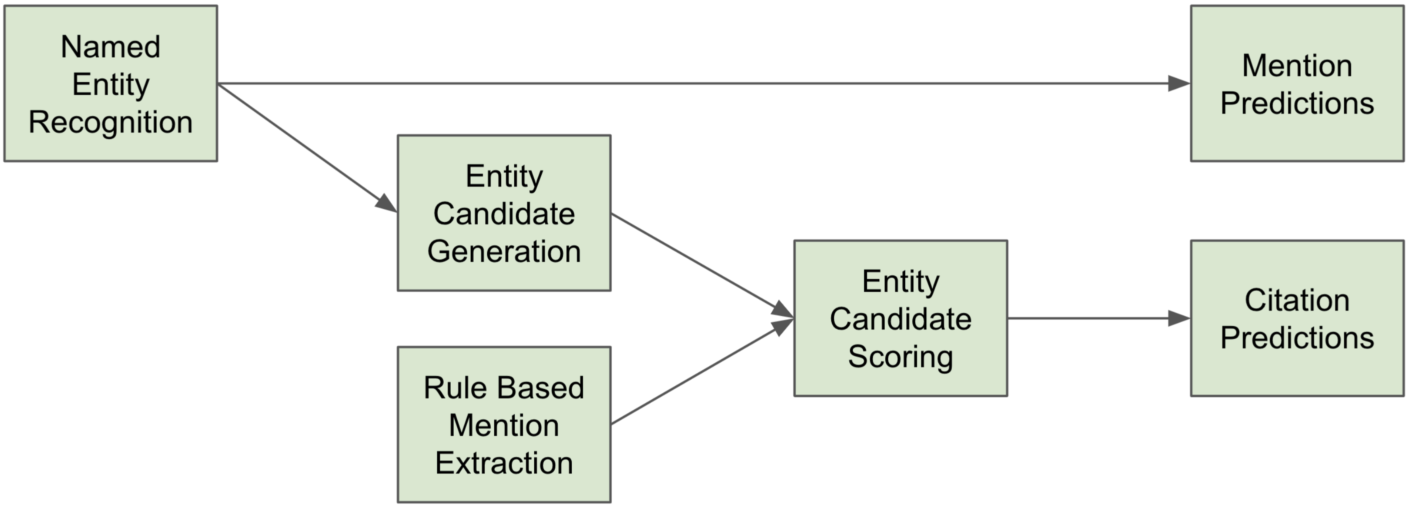


Figure 6.1: A high-level overview of our appraoch to dataset mention detection and linking.

We provide a high-level overview of our approach in Figure 6.1. First, we use an NER model to predict dataset mentions. For each mention, we generate a list of candidate datasets from the knowledge base. We 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. We then use a binary classifier to predict which of these candidates is a correct dataset extraction.

Next, we describe each of the sub-components in more detail.

Mention and Candidate Generation

We first constructed a set of rule based candidate citations by exact string matching mentions and dataset names from the provided knowledge base. We found this to have high recall and low precision, both on the provided development fold and our own development fold that we created. High recall and low precision was the desired outcome for this candidate generation step. However, after our test submission, it became clear that there were many datasets in the actual test set that did not have mentions in the provided knowledge base. If the provided development fold had been representative of the test set (rather than the train set) in terms of what datasets were mentioned in it, we could have discovered this issue sooner. In this case, it would have been more representative of the test set if it included more datasets that did not have example mentions in the provided knowledge base. The importance of reliable evaluation and training data is discussed further in Chapter 12.

To address this limitation, we developed an NER model to predict additional dataset mentions. For NER, we use a bi-LSTM model with a CRF decoding layer, similar to Peters et al., 2018, and implemented using the AllenNLP framework[2]. In order to train the NER model, we automatically generate mention labels by string matching mentions in the provided annotations against the full text of a paper. This results in noisy labeled data, because it was not possible to find all correct mentions this way (e.g., some dataset mentions were not annotated), and the same string can appear multiple times in the paper, while only some are correct examples of dataset usage.

We limit the percentage of negative examples (i.e., sentences with no mentions) used in training to 50%, and use 40 words as the maximum sentence length. We use 50-dimensional Glove word embeddings (Pennington et al., 2014), 16-dimensional character embeddings with 64 CNN filters of sizes (2, 3, 4). The CNN character encoder outputs 128-dimensional vectors. We optimize model parameters using ADAM (Kingma and Ba, 2014) with a learning rate of 0.001. Training the model took approximately 12 hours on a single GPU.

In order to generate linking candidates for the NER mentions, we score each candidate dataset based on TF-IDF weighted token overlap between the mention text and the dataset title. For a given mention, many dataset titles can have a non-zero overlap score, so we take the top 30 scoring candidates for each mention as the linking candidates for that mention.

Candidate Linking

The linking model takes as input a dataset mention, its context, and one of the candidate datasets in the knowledge base, and outputs a binary label. We use a gradient boosted trees classifier using the XGBoost[3] implementation. The model takes as input the following features:

- prior probability of entity, estimated based on number of occurrences in the training set (float between 0 and 1)
- prior probability of entity given mention, estimated based on number of occurrences in the training set (float between 0 and 1)
- prior probability of mention given entity, estimated based on number of occurrences in the training set (float between 0 and 1)
- whether the same year appears both in the mention context and in the dataset title (binary)
- mention length (int)
- mention sentence length (int)
- whether the mention is an acronym, computed by checking if it is one token that is all upper case (binary)
- estimated section title of the mention, computed by searching backwards from the mention for the nearest section header (binary one-hot)
- count of overlapping words between the mention context and dataset keywords provided in the knowledge base (int)

We note that it is possible to predict zero, one or multiple dataset IDs for the same mention, and each dataset candidate is scored independently.

We performed a randomized hyperparameter search with 100 iterations over the following hyperparameters and ranges and used a learning rate of 10^–1:

- `max_depth` : `range(2, 8)`
- `n_estimators` : `range(1, 50)`
- `colsample_by_tree` : `numpy.linspace(0.1, 0.5, 5)`
- `min_child_weight` : `range(5, 11)`

Each model took a negligible amount of time to train, and the entire hyperparameter search took a few minutes to train on a machine with 8 CPUs.

Research Area Prediction

Data

The second task of the competition is to predict research areas of a paper. The task does not specify the set of research areas of interest, nor is training data provided for the task. After manual inspection of a subset of the papers in the provided test set, the SAGE taxonomy of research, and the Microsoft Academic Graph (MAG) (Shen et al., 2018), we decided to use a subset of the fields of study in MAG as labels. In particular, we included all fields related to social science or papers from the provided training corpus. However, since the abstract and full text of papers are not provided in MAG, we only use the paper titles for training our model. The training data we ended up with included approximately 75K paper titles along with their fields of study as specified in two levels of the MAG hierarchy. We held out about 10% of the titles for development data. The coarse level (L0) has 7 fields while the more granular one (L1) has 32. Fields associated with less than 100 papers were excluded.

Methods

For each level, we trained a bi-directional LSTM which reads the paper title and predicts one of the fields in this level. We additionally incorporate ELMo embeddings (Peters et al., 2018) to improve performance. In the final submission, we always predict the most likely field from the L0 classifier, and only report the most likely field from the L1 classifier if its prediction exceeds a score of 0.4. It takes approximately 1.5 and 3.5 hours for the L0 and L1 classifiers to converge, respectively.

Research Method Extraction

Data

The third task in the competition is to extract the scientific methods used in the research paper. Since no training data was provided, we started by inspecting a subset of the provided papers to get a better understanding of what kind of methods are used in social science and how they are referred to within papers. This limitation, and the difficulty of working on an undefined task, is discussed in Chapter 2.

Methods

Based on the inspection, we designed regular expressions which capture common contextual patterns as well as the list of provided SAGE methods. In order to score candidates, we used a background corpus to estimate the salience of candidate methods in a paper. Two additional strategies were attempted but proved unsuccessful: a weakly-supervised model for named entity recognition, and using open information extraction (openIE) to further generalize the list of candidate methods.

Results

Dataset Extraction and Linking

First, we report the results of our NER model in Table 6.1. Since it is easy for the model to memorize the dataset mentions seen at training time, we created disjoint train, development, and test sets based on the paper–dataset annotations provided for the competition. In particular, we sort datasets by the number of papers they appear in, then process one dataset at a time. For each dataset, we choose one of the train, development, or test splits at random and add the dataset to it, along with all papers which mention that dataset. When there is a conflict, (e.g., a paper *p* has already been added to the train split when processing an earlier dataset d_1, but it is also associated with a later dataset d_2), the later dataset d_2 along with all papers associated with it are added to the same split as d_1. For any further conflicts, we prefer to put papers in the development split over the train split, and the test split over the development split.

We also experimented with adding ELMo embeddings (Peters et al., 2018), but it significantly slowed down training and decoding which would have disqualified our submission due to the runtime requirements of the competition. As a result, we decided not to include ELMo embeddings in our final model. If the requirements for the competition had permitted the use of a GPU at evaluation time, neural network based embeddings like ELMo could be leveraged.

	prec.	recall	F1
dev set	53.4	50.3	51.8
test set	50.7	41.8	45.8

Table 6.1: NER precision, recall and F1 performance (%) on the development and test sets.

	prec.	recall	F1
baseline	28.7	58.0	38.4
+ p(d|m), p(m|d)	39.6	42.0	40.7
+ year matching	35.1	57.0	43.5
+ aggregated mentions, tuning, and other features	72.5	45.0	55.5
+ dev set examples	77.0	47.0	58.3
+ NER mentions	56.3	62.0	59.0

Table 6.2: End-to-end precision, recall and F1 performance (%) for citation prediction on the development set provided in phase 1 of the competition.

	prec.	recall	F1
phase 1 holdout	35.7	19.6	25.3
phase 2 holdout	39.6	18.8	25.5

Table 6.3: End-to-end precision, recall, and F1 performance (%) for dataset prediction on the phase 1 and phase 2 holdout sets. Note that the phase 1 holdout results are for citation prediction, while the phase 2 holdout results are for mention prediction.

We report the end-to-end performance of our approach (on the development set provided by the organizers in the first phase) in Table 6.2. This is the performance after using the linking classifier to predict which candidate mention–dataset pairs are correct extractions. We note that the development set provided in phase 1 ended up having significantly more overlap with the training data than the actual test set did. As a result, the numbers reported in Table 6.2 are not indicative of test set performance.
End to end performance from our phase 2 submission can be seen in Table 6.3. This performance is reflective of our focus on the linking component of this task. Aside from the competition development set, we also used a random portion of the training set as an additional development set. The initial model only uses a dataset frequency feature, which gives a baseline performance of 38.4 F1. Adding p(d | m) and p(m | d), which are the probability of entity given mention and probability of mention given entity improves the performance ($\Delta = 2.3$ F1). Year matching helps disambiguate

between different datasets in the same series, which was found to be a major source of errors in earlier models ($\Delta = 2.8$ F1). Aggregating mentions for a given dataset, adding mention and sentence length features, adding an is acronym feature, and further hyper-parameter tuning improve the results ($\Delta = 12.5$ F1). Adding examples in the development set while training the model results in further improvements ($\Delta = 2.8$ F1). Finally, adding the NER-based mentions significantly improves recall at the cost of lower precision, with a positive net effect on F1 score ($\Delta = 0.7$ F1).

Two clear limitations of our model are its difficulty in generalizing to unseen datasets, and its inability to effectively distinguish between datasets that are used in a publication and datasets that are merely referenced. These limitations are the main causes of the low recall (due to difficulty generalizing to unseen datasets) and low precision (due to difficulty distinguishing between used datasets and referenced datasets). An interesting approach to improving recall is presented in Chapter 7, and could potentially be leveraged in future work.

Research Area Prediction

To select a model, we performed a 100 trial random search across model hyper-parameters, evaluated on a held out development set of papers from the Microsoft Academic Graph. Our final model contained 512 hidden dimensions, 2 layers and 0.5 dropout prior to classification. The top performing classifier achieved 84.4% accuracy on our development set on L0 fields, and 65.2% accuracy on our development set on L1 fields. The main limitation of using MAG for this problem is that our model cannot find new fields of research, and is limited to those provided by MAG. Additionally, our method performs classification based only on the titles of papers, while there are other pieces of information about the paper that would be useful for classifying the field of research. Other resources that could have been used to help with this task are presented in Chapters 7 and 8.

Research Method Extraction

We evaluated performance by manually evaluating the output of our extractor for a subset of 50 papers from the provided test set to compute precision. Since evaluating recall requires a careful annotation, we resorted to using yield as an alternative metric. Our final submission for method extraction has 95% precision and yield of 1.5 methods per paper on the manually inspected subset of papers. Similarly to research area prediction, the main limiation here is the difficulty our model has finding new methods, as it is limited to the SAGE ontology and a few hand-crafted patterns. One potential way to alleviate this issue is to leverage external resources, as presented in Chapter 8.

Conclusion

This report summarizes the AI2 submission at the RCC competition. We identify dataset mentions by combining the predictions of an NER model and a rule-based system, use TF-IDF to identify candidates for a given mention, and use a gradient boosted trees classifier to predict a binary label for each candidate mention–dataset pair. To identify research fields of a paper, we train two multi-class classifiers, one for each of the top two levels in the MAG hierarchy for fields of study. Finally, to extract research methods, we use a rule-based system utilizing a dictionary and common patterns, followed by a scoring function which takes into account the prominence of a candidate in foreground and background corpora.

We now provide some possible directions of improvement for each component of our submission. For dataset extraction, the most promising avenue of improvement is to improve the NER model, and the most promising avenue to improve the NER model is to collect less noisy data. We effectively have distantly supervised training data for the NER model, and the first thing to try would be directly annotating papers with dataset mentions to provide a clearer signal for the NER model. As mentioned previously and discussed in Chapters 5 and 8, the dataset mentions provided are not located in the text, and are simply extracted strings. Given these strings, labels in the actual text can be created by searching for the provided string. However, this is a noisy process, as the string may occur multiple times in the document, and all occurrences may or may not be correct dataset mentions. This is especially problematic when the string is a common word (e.g. "time"). Therefore, directly annotating the strings in the full text (i.e. providing character offsets for the strings) would help to reduce the noise in the NER data. For
research area prediction, it would help to include signals beyond just the paper title for predicting the field of study. The difficulty here is finding labeled training data that includes richer signals like abstract text and paper keywords. For method prediction, exploring the use of open information extraction is a potential avenue of future research. Additionally, it would be helpful to clarify what exactly is meant by a method, as it is currently unclear what a successful method extraction looks like.
The main lesson learned is that, when presented with noisy, distantly supervised, real-world data, to produce a production-quality system, it becomes very important to (1) have a high-confidence evaluation dataset, and (2) look for other data sources that are similar enough to the task at hand to be useful. Taking steps towards both of these objectives are promising avenues of future work.

As discussed in Chapter 1 and throughout, the dataset extraction discussed in this report is intended to be part of a broader effort to create infrastructure and tools that will aid in the discovery and usage of datasets in social science research. This is critical to enable reproducibliity, collaboration, and effective use of data. We look forward to seeing what comes out of this project as a whole, and how AI techniques can be leveraged to have positive impact in the social sciences.

Acknowledgments

We would like to thank the competition organizers for their tireless efforts in preparing the data, answering all our questions, doing the evaluations, and providing feedback. We also would like to thank Zhihong (Iris) Shen for helping us use the MAG data.

References

Diederik P. Kingma and Jimmy Ba. 2014. Adam:A method for stochastic optimization.CoRR,abs/1412.6980.

Jeffrey Pennington, Richard Socher, and Christo-pher D. Manning. 2014. Glove: Global vectors forword representation. InEMNLP.

Matthew E. Peters, Mark Neumann, Mohit Iyyer,Matt Gardner, Christopher Clark, Kenton Lee, andLuke S. Zettlemoyer. 2018.
Deep contextualizedword representations. InNAACL 2018.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018.A web-scale system for scientific knowledge explo-ration. InACL

Footnotes

1: www.semanticscholar.org

2: https://github.com/allenai/allennlp/blob/master/allennlp/models/crf_tagger.py

3: https://xgboost.readthedocs.io/en/latest/

4: https://github.com/allenai/coleridge-rich-context-ai2

Appendix

The code for the submission can be found on GitHub[4]. There is a README with additional documentation at this github repo.

Chapter 7 - Finding datasets in publications: The KAIST approach

author:
- |
Haritz Puerto-San-Roman
IR & NLP Lab
KAIST
Daejeon, South Korea
haritzpuerto94 @kaist.ac.kr
Giwon Hong
IR & NLP Lab
KAIST
Daejeon, South Korea
gch02518 @kaist.ac.kr
Minh-Son Cao
IR & NLP Lab
KAIST
Daejeon, South Korea
minhson @kaist.ac.kr
Sung-Hyon Myaeng
IR & NLP Lab
KAIST
Daejeon, South Korea
myaeng@kaist.ac.kr
bibliography:
- ‘bibliography.bib’

title: ‘Finding datasets in publications: The KAIST approach. Text Mining Using Question Answering’

Non-technical Overview

The KAIST’s approach for retrieving datasets is to generate questions about datasets like *what is the dataset used in this publication?* and use a machine-learning system that can read a publication and a question, and give the answer to it. This machine-learning system retrieves a list of candidate answers among which one of them should be the name of the dataset. To remove those wrong candidate answers, they proposed to filter them out by their entity types. For example, if the entity type of a candidate answer is *organization*, it is likely to be a dataset name because datasets are created by organizations.

For research field retrieval, they proposed to compare publications with Wikipedia articles to discover the research fields. First, they crawled Wikipedia articles that correspond to the list of research fields. Then, they retrieved the research fields of the publications by measuring the similarity between the papers and the crawled Wikipedia documents. For example, they crawled the Wikipedia article *economic history* which corresponds to the research field *economic history*. If the similarity between a publication and the article *economic history* is high enough, it is determined that the publication belongs to the research field *economic history*. They proposed to use TF-IDF as similarity measure, which is based on term frequency and document frequency, but others could be applied too.

For the research methods retrieval, they modeled the task as a named-entity recognition problem. They considered research methods as named entities, real-world objects that can be denoted with a proper name, and trained a machine learning model to identify and retrieve them.

Literature Review

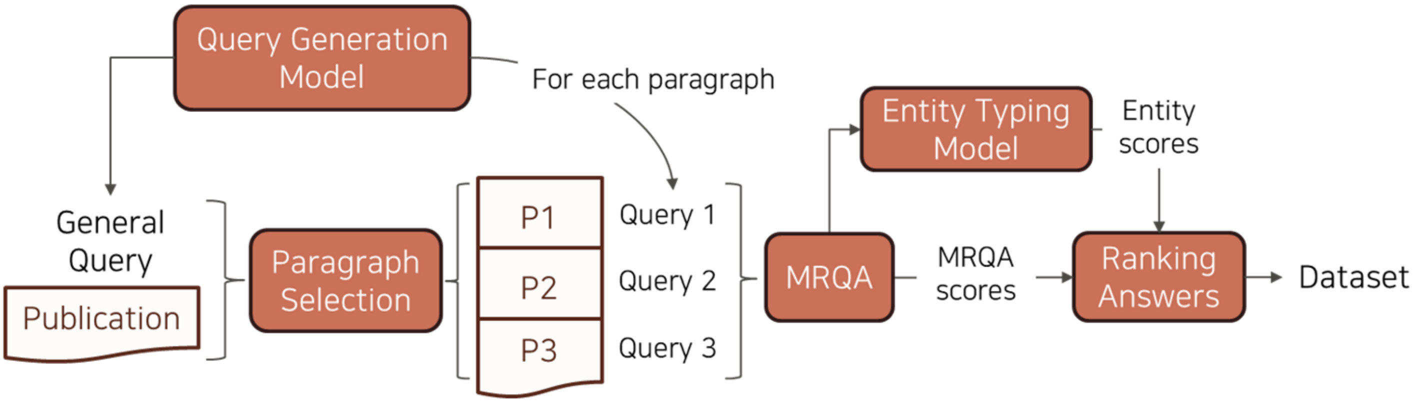
Although *Information Retrieval* is a well-established research field, only a few attempts have focused on the task of dataset extraction form publications. [@ghavimi2016identifying] tackled this problem using heuristics and dictionaries but encountered several problems. Firstly, they gave too much weight to acronyms. For example, *NYPD (New York Police Department)* is detected as a dataset name.
Secondly, they gave too much weight to the year of publication of the datasets because they assumed that dataset names are usually followed by the year of publication. However, this may only apply to Social Sciences publications. For example, Computer Science datasets do not appear followed by the publication year so this heuristic cannot detect all possible types of dataset mentions.

What did you do

In this section, a detailed explanation of the used models for dataset names, research field, and research methods retrieval is provided.

Datasets Retrieval

The followed approach to retrieve dataset names is based on Machine Reading Question Answering (MRQA). First, given a publication, a list of candidate paragraphs in which the dataset is mentioned is selected. Then, using a query generation module, a specific query for each paragraph is created. After that, each pair of paragraph-query is input into the MRQA model. This model creates a list of candidate answers that is further processed using a feed-forward neural network. This network takes as input pairs of candidate answers and their entity types. The types of the answer candidates are obtained using an entity typing module. The output of this feed-forward neural network is the final list of dataset names found in the publication. Figure 7.1 shows an overview of the pipelined system. In the following subsections, the MRQA, query generation, and entity typing models are explained in detail.



image

Figure 7.1: Overall architecture for dataset retrieval

Document QA

MRQA models are neural networks that find answers for given queries according to a text. These answers must appear explicitly in the text. Since the dataset retrieval task is about finding explicit dataset mentions from publications, MRQA models are suitable for this task.

The MRQA model used in this work is Document QA [@clark2017simple]. It uses Bi-GRU, bi-attention, and self-attention mechanism. In addition, Document QA performs a paragraph selection that pre-filters and selects the *k* most relevant paragraphs using TF-IDF similarity between the query and paragraphs of the text. The model was trained on SQuAD v1.1, a common dataset to train MRQA models [@rajpurkar2016squad]. For details about the implementation and computing resources to train it, they refer to the original publication.

The KAIST team had the hypothesis that MRQA models do not need the full publication to find datasets. Rather, the MRQA models only need to process the paragraph where the answer appears. Among the literature of MRQA, the model used in this work stands out because of its paragraph selection stage. Using this model, it is possible to select a list of candidate paragraphs where the answer may appear, and then use the MRQA model to process them and retrieve the datasets.

Query Generation Module

Queries that are suitable for finding datasets are required to utilize an MRQA model for the dataset retrieval task. However, defining a general query for retrieving datasets is not trivial, since the dataset mentions appear in various forms like surveys, datasets, or studies. Therefore, they devised a query generation module to generate multiple specific queries instead of a single general query.

To create a list of important query terms that the queries should include, they used a query generation model proposed by [@yuan2017machine] that creates a query given a text and an answer. All the queries generated by this model are too specific and cannot be used for other publications or other dataset names. However, they can be utilized to create a list of query terms to generate more specific queries for other publications. To create this list, they extracted query terms that are frequent in the list of queries and at the same time are not frequent in sentences that do not include a mention to a dataset. Because of this, these query terms have the discriminative power to retrieve dataset mentions since 1) queries are generated to extract mentions and 2) the query terms do not appear in the sentences without dataset mentions.
This list is used two times in their pipelined system. First, concatenating the first *k* query terms a general query is built. This query is employed by the paragraph selection stage of Document QA, as shown in Figure 7.1. Then, the query generation module generates specific queries for each paragraph concatenating the query terms that appear in the paragraph and on the list.

Entity Typing Model

Ultra-Fine Entity Typing [@Choi:2018:ACL] can predict a set of free-from phrases like *criminal* or *skyscraper* given a sentence with an entity mention. For example, in the sentence: *Bob robbed John and he was arrested shortly afterward*, Bob is of type *criminal*. In the task of the present book, candidate answers proposed by the MRQA model and the sentence in which they appear are input into Ultra-Fine Entity Typing. This system can predict 10k different entity types among which *dataset* is included. However, after a few experiments, they observed that most of the entity types obtained from the dataset names are not *dataset* but *organization*, *agency*, and similar types. This is due to the fact that datasets are usually created by organizations and thus, they include the name of the organization in the name of the dataset. Since these entity types are consistent, it is possible to use them as a feature for their candidate answer classifier. In this work, the KAIST team used the pre-trained model that was released with the original publication. For details about the implementation and computing resources to train it, they refer to the original publication.

Candidate Answer Classifier

Using the score given by the MRQA model for each candidate answer and the entity types given by the Entity Typing model for each candidate answer, a neural network classifier that filters the wrong candidate answers provided by the MRQA model is used. The intuition of this classifier is that a candidate answer with a high score given by the MRQA model and whose entity type is *organization* or something similar is highly likely to be a correct dataset name. Due to this pattern, they were able to create a neural network classifier to filter out candidate answers.

The classifier has the following architecture:

1. Input size: 10332 (10331 labels from Ultra-Fine Entity Typing and one from the Document QA score)
2. 1 hidden layer with 50 neurons
3. Output size: 2

The training set consists of 25172 examples and the test set of 6293 examples obtained from the training set provided by the competition. To train the network, Adam optimizer was used with cross entropy as the loss function.

Research Fields Retrieval

Their approach to obtain research fields is based on comparing the publications with Wikipedia articles using TF-IDF similarity. First, using the list of research fields provided by the competition, a set of Wikipedia articles about different research fields was obtained using the Python library MediaWiki. The list provided by the competition has three levels of hierarchy as shown in the example (Figure 7.2). The leaf nodes of that hierarchy were searched in Wikipedia to retrieve specific research fields instead of general ones. For example, they were aiming to retrieve *Neurosurgery* instead of *Medicine*. Then, using Scikit-learn [@scikit-learn], a TF-IDF matrix of all the publications and Wikipedia articles of research fields were computed. The research field and all its superior nodes in the hierarchy associated with the Wikipedia article most similar to the publication were returned along with the similarity in the range [0,1]. The overall architecture can be seen in Figure 7.3.

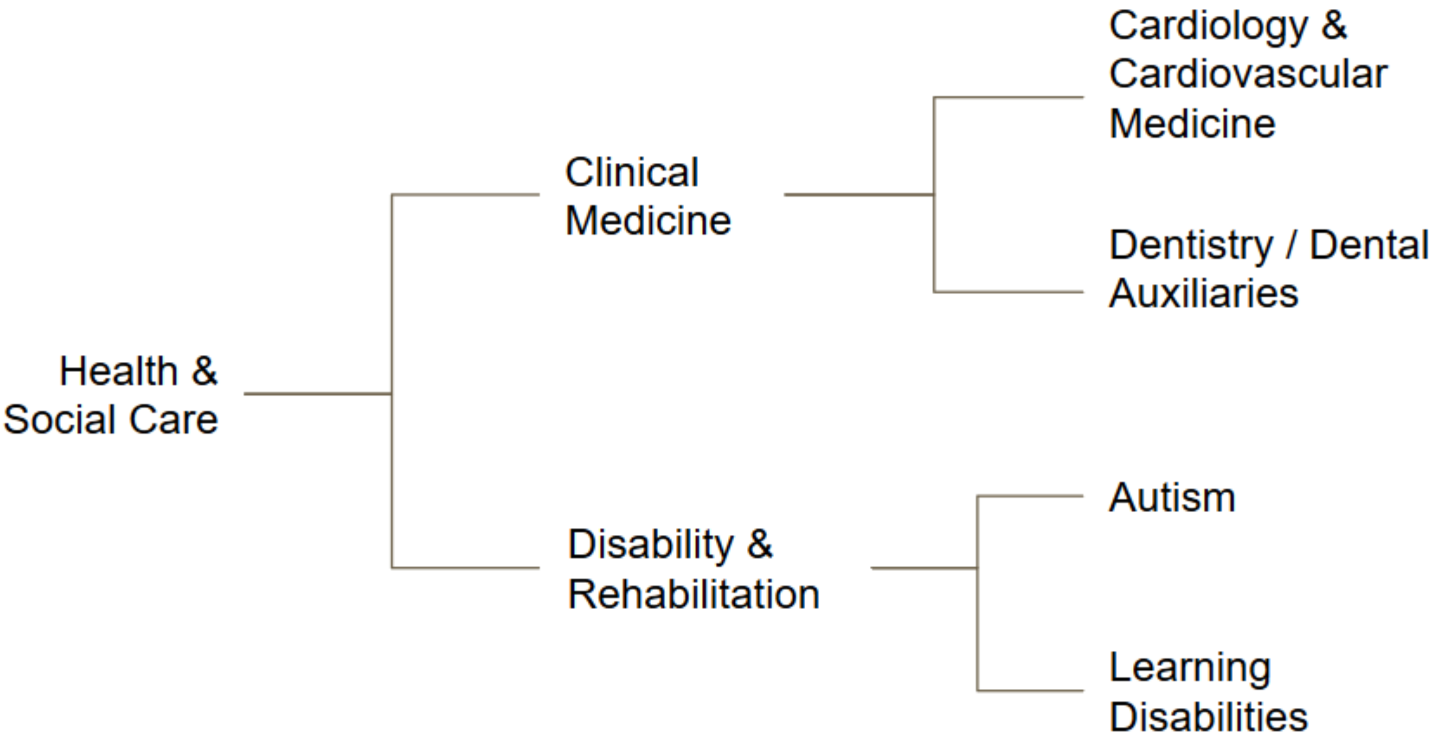

Health & Social Care
Clinical Medicine
Cardiology & Cardiovascular Medicine
Dentistry / Dental Auxiliaries
Disability & Rehabilitation
Autism
Learning Disabilities

Figure 7.2: Research fields hierarchy

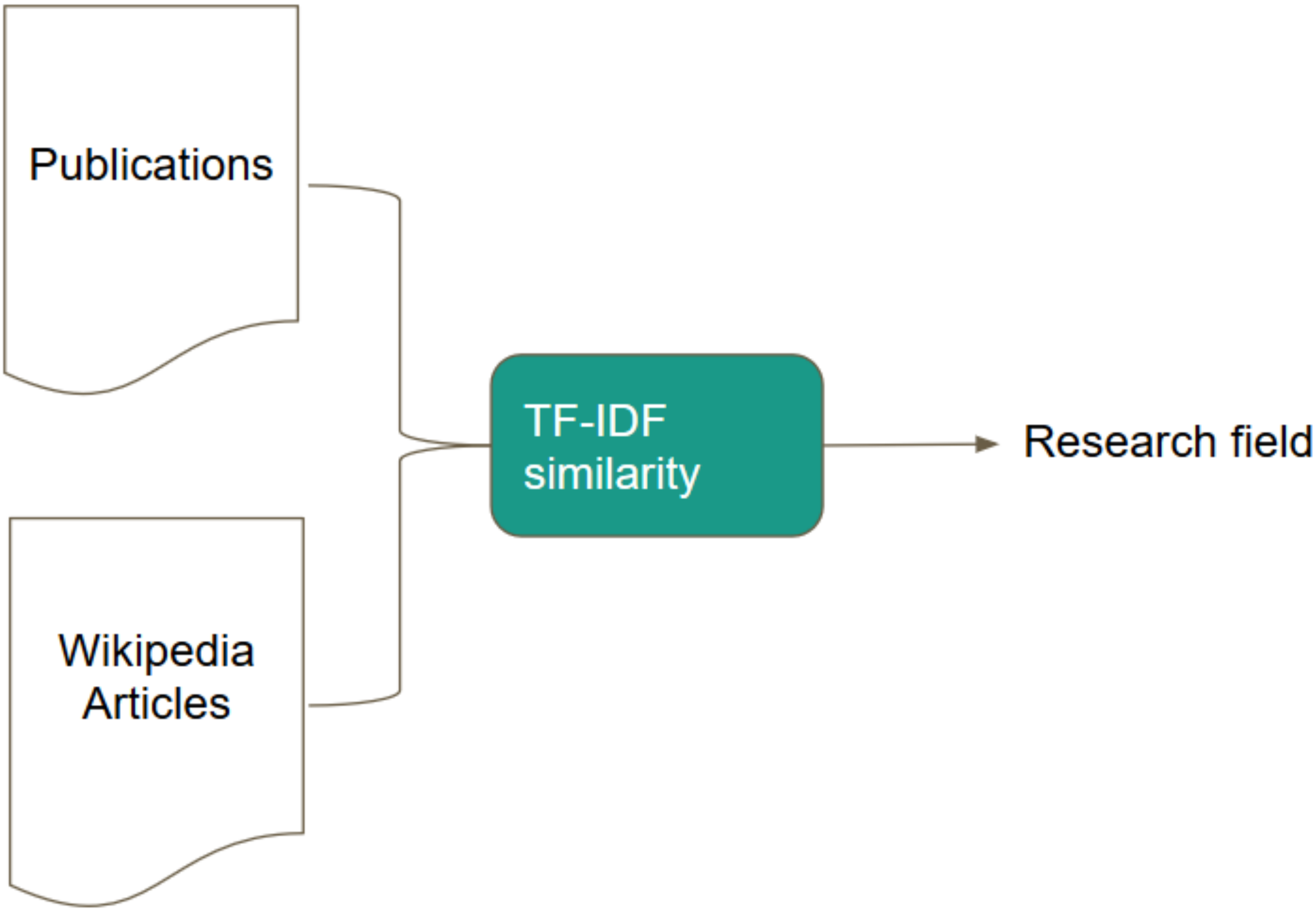


Figure 7.3: Overall architecture for research fields retrieval

Research Methods Retrieval

For the research methods retrieval task, they modeled it as a named-entity recognition (NER) problem. Research methods are considered to be named entities and because of this, they can be tagged as research method label (RS) instead of common NER labels such as *location*, and *people*. Figure 7.4 shows the main architecture of the model proposed by [@lample2016neural] and used in this task.

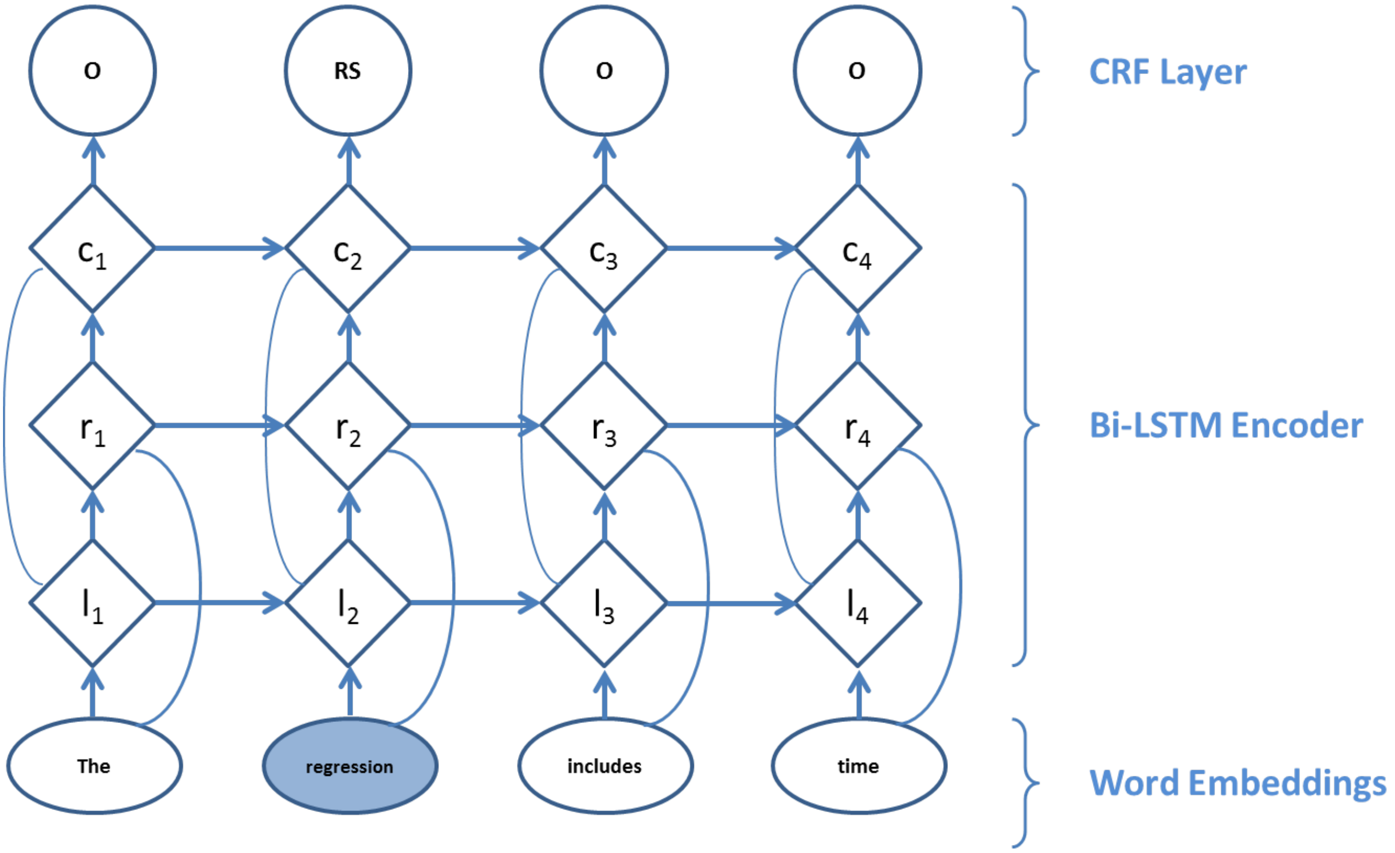


Figure 7.4: Paragraph selection for DocQA in research method retrieval

The representation of a word using the model is obtained considering its context. The KAIST team had the assumption that research methods have dependencies and constraints with words that appear in their surrounding context. Therefore, the conditional random field [@lafferty2001conditional] layer in this model is suitable for detecting research methods by jointly tagging the whole sentence, instead of independently tagging each word.

For this task, the research method phrases that appeared in the training set were marked using as reference the list of research methods provided by the competition. Then, the training set was represented in CoNLL 2003 format [@tjong2003introduction], using IOB tag (Inside, Outside, Beginning) to train the model. Every token was labeled as B-RS if the token is the beginning of a research method, I-RS if it is inside a research method but not the first token, or O otherwise. Training the model took approximately one day using a CPU AMD Ryzen 7 2700, a GPU Nvidia GeForce GTX 1050 Ti, and 8GB RAM.

What worked and what didn’t

The KAIST team tried different ideas to extract dataset names. First, they tried to extract the dataset names using hand-crafted queries in the MRQA model. However, they noticed that these manually generated queries do not have sufficient discriminative power. Therefore, they tried to generate a general query with enough discriminative power to retrieve datasets’ names. To this end, they converted the sentences containing the dataset into queries and then clustered the converted queries to get general queries. However, they found that each of the resulting clusters did not reflect the semantics of the desired general queries. Hence, they had to create specific queries for each publication as explained in the previous section. These specific queries helped to increase the recall but at the same time affected negatively to the precision. The use of entity typing worked well to remove the wrong candidate answers proposed by the MRQA model. Thanks to this entity-type filtering, they were able to improve the recall using the query generation module without sacrificing the precision.

They also tried to use the section names as a feature for the paragraph selection module of Document QA. However, the use of section names degraded the overall performance. In their analysis, they stated that the heuristics that they used to extract them generated noise that affected the performance. For example, *total*, *variable*, and *funding* were detected as section names but clearly, they are not.

Finally, their idea to compare publications with Wikipedia articles to retrieve research fields yielded successful results. But on the other hand, their first idea to retrieve research methods was not successful. It was based on identifying the context words of the research methods by using the frequency of those words. The reason for the bad performance was due to the lack of discriminative power of the most common words that co-occur with the research methods. Therefore, they tried to model it as a NER problem, where they consider each research method that appeared in a publication as a named-entity. By modeling the problem in this way, they could use existing NER models to extract research methods from publications. However, this approach did not achieve satisfactory results either.

Summary of your results and caveats

Due to the difficulty of performing a quantitative analysis on a not extensively labeled dataset, a qualitative analysis was made. Several random publications were chosen and manually labeled to check the quality of the model and discover the strong and weak points.

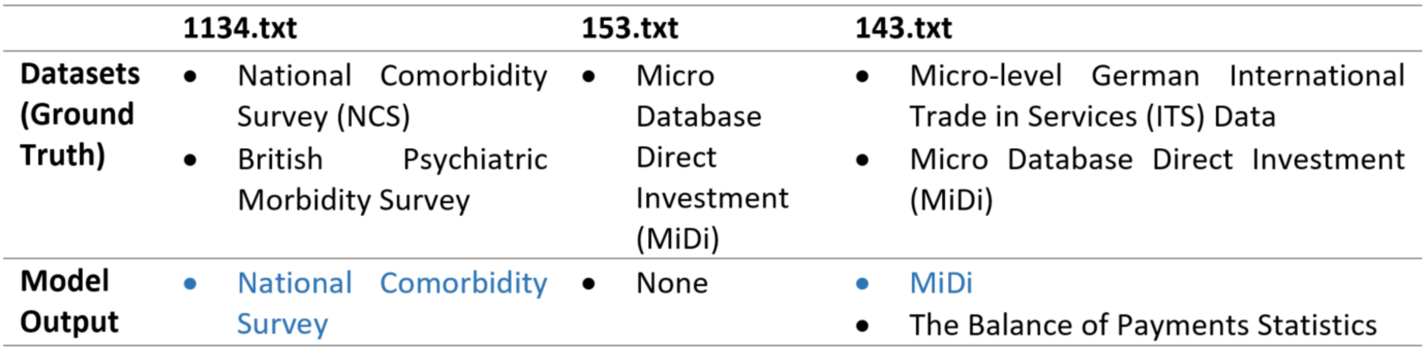
Datasets Retrieval

To analyze the effects of the query generation module and entity typing module, they performed analyses on 100 publications from the dev set of the first phase using three different settings:

1. Document QA only
2. Document QA + Query Generation Module
3. Document QA + Query Generation Module + Entity Typing Module

Document QA only

Figure 7.5 shows the results from three publications from the list of analyzed publications using *Document QA only*. Compared to the other settings, *Document QA only* setting retrieves high-quality answers (dataset mentions). However, the number of retrieved answers is notably small. For example, the result from *153.txt* publication is empty as shown in Figure 7.5. In fact, using this setting the model can only retrieve 260 answers (predictions) from the list of analyzed publications.



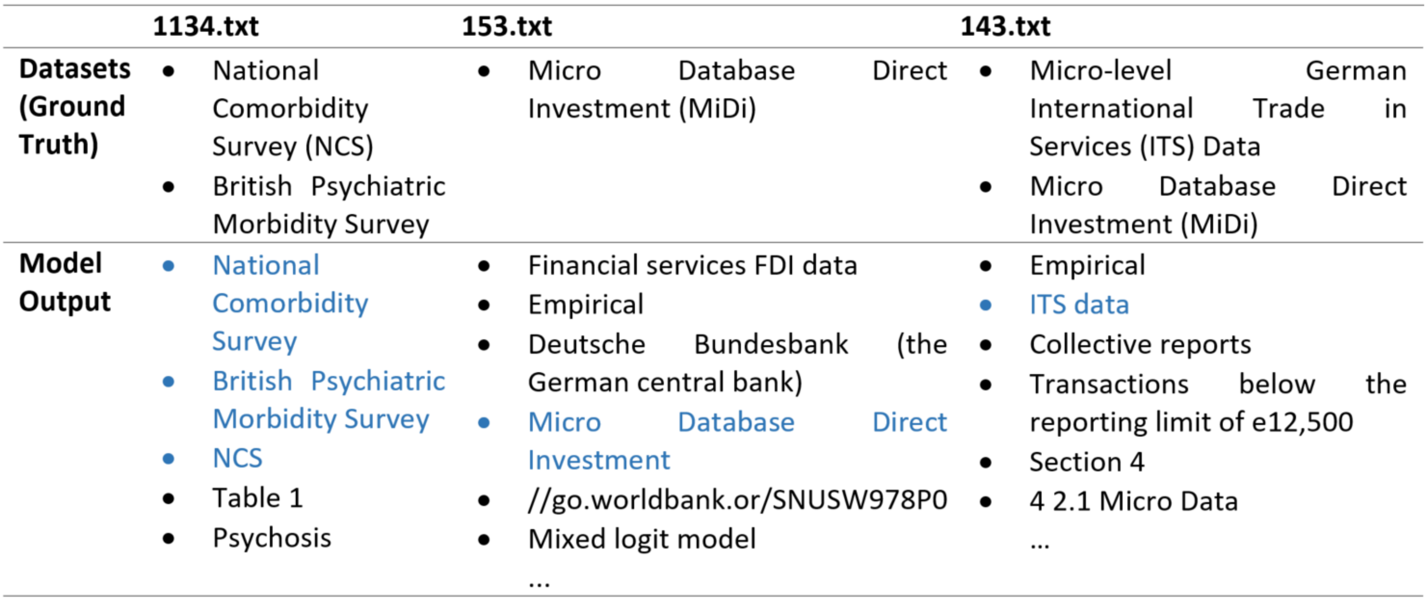
	1134.txt	**153.txt**	**143.txt**
Datasets (Ground Truth)	• National Comorbidity Survey (NCS) • British Psychiatric Morbidity Survey	• Micro Database Direct Investment (MiDi)	• Micro-level German International Trade in Services (ITS) Data • Micro Database Direct Investment (MiDi)
Model Output	• National Comorbidity Survey	• None	• MiDi • The Balance of Payments Statistics

Figure 7.5: Results from Document QA only. Right answers from the model in blue.

These results were expected due to the difficulty of defining general queries as explained in section *Question Generation Module*. Without a query generation module, it is hard to make a representative enough query to retrieve various forms and types of the dataset mentions.

Document QA + Query Generation Module

Figure 7.6 shows the results from three publications from the list of analyzed datasets using Document QA and the Query Generation Module. Because of the addition of the Query Generation Module, a larger number of answers were retrieved. For example, the result from *153.txt* publication contains several answers including the right one, *Micro Database Direct Investment*. Therefore, this and the retrieval of more than 2,000 answers from the list of analyzed datasets proves that the Query Generation Module improves recall of the entire dataset retrieval model. On the other hand, compared to the *Document QA only* setting, there is a considerable amount of wrong candidate answers. For instance, in Figure 7.6, *empirical*, *Table 1*, and *Section 4* are not dataset mentions.



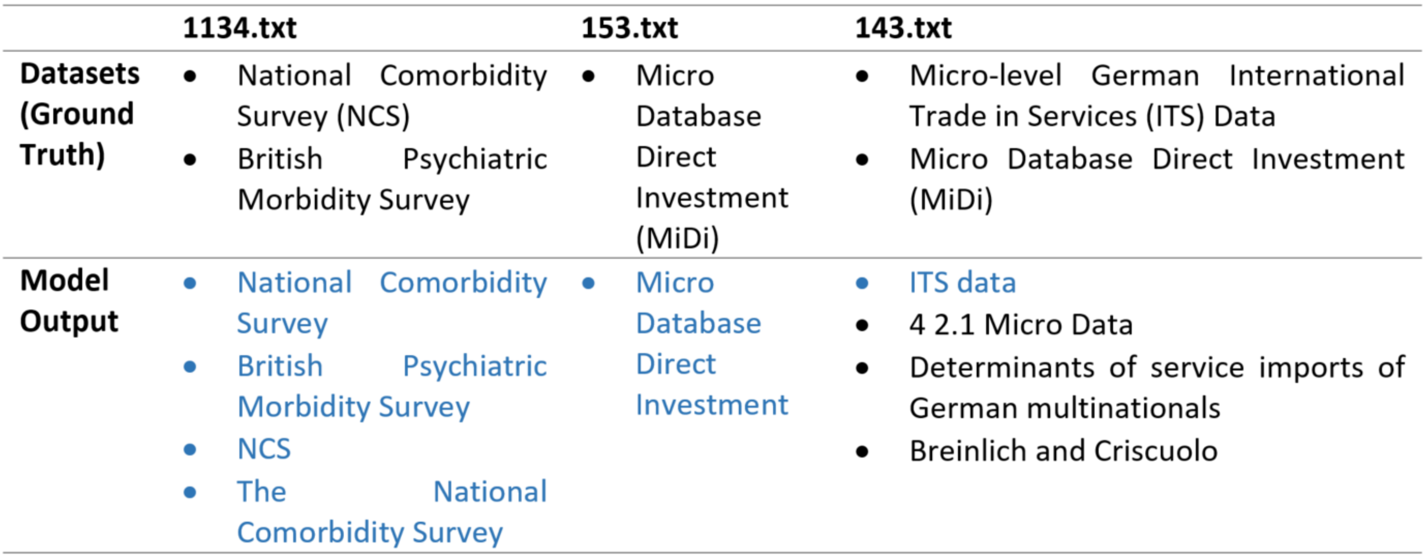
	1134.txt	153.txt	143.txt
Datasets (Ground Truth)	• National Comorbidity Survey (NCS) • British Psychiatric Morbidity Survey	• Micro Database Direct Investment (MiDi)	• Micro-level German International Trade in Services (ITS) Data • Micro Database Direct Investment (MiDi)
Model Output	• National Comorbidity Survey • British Psychiatric Morbidity Survey • NCS • Table 1 • Psychosis	• Financial services FDI data • Empirical • Deutsche Bundesbank (the German central bank) • Micro Database Direct Investment • //go.worldbank.or/SNUSW978P0 • Mixed logit model ...	• Empirical • ITS data • Collective reports • Transactions below the reporting limit of e12,500 • Section 4 • 4 2.1 Micro Data ...

Figure 7.6: Results from Document QA + query generation module. Right answers from the model in blue.

They believe that the reason for this noise is that some query terms may cause the retrieval of wrong answers. For example, the query term *study* can help to retrieve dataset mentions such as *ANES 1952 Time Series Study*. However, this term can also retrieve wrong answers such as *empirical study*. These types of query terms are still needed to retrieve various forms and types of dataset mentions but clearly generate noise.

Document QA + Query Generation Module + Entity Typing Module

Figure 7.7 shows the results of the analyzed publications using Document QA, the Query Generation Module, and the Entity Typing Module. Although the Entity Typing Module might not remove all the wrong answers caused by the Query Generation Module such as *4 2.1 Micro Data*, most of them are successfully removed and thus, the overall precision is improved. Using this setting the model retrieves 526 answers (predictions) from 100 publications from the dev set of the first phase of the competition.



	1134.txt	153.txt	143.txt
Datasets (Ground Truth)	• National Comorbidity Survey (NCS) • British Psychiatric Morbidity Survey	• Micro Database Direct Investment (MiDi)	• Micro-level German International Trade in Services (ITS) Data • Micro Database Direct Investment (MiDi)
Model Output	• National Comorbidity Survey • British Psychiatric Morbidity Survey • NCS • The National Comorbidity Survey	• Micro Database Direct Investment	• ITS data • 4 2.1 Micro Data • Determinants of service imports of German multinationals • Breinlich and Criscuolo

Figure 7.7: Results from Document QA + query generation module + entity typing module. Right answers from the model in blue.

Research Fields Retrieval

They randomly selected 20 publications from the training set of the first phase. They can test their model using the training set because the model does not require any training phase. The model was able to predict research fields correctly for 11 of those publications. The strongest point is that the model is able to predict research fields that are significantly specific such as *Home health nursing management*. Among the weak points of the model, it has problems when two research fields are similar or share subtopics. Moreover, sometimes it fails due to the fact that it tries to retrieve excessively specific fields while more general ones would be suitable.

Research Methods Retrieval

20 random publications were selected and labeled from the training set of the second phase and the results are not satisfactory. The model is able to find proper research methods for 12 publications out of 20. For example, the model successfully detects the research method of the publication with id 15359, *Factor analysis*. However, the results contain a significant amount of noise. For example, the model retrieves for the document with id 10751 several wrong answers like *Reviews describe*, *Composite materials*, and *Detailed databases*.

Lessons learned and what would you do differently

After the completion of this project, the KAIST team realized that some steps could have been in a different way and would have led to better results. For example, they focused a lot on the model creation, but they could have spent more time on the analysis of the dataset to extract all its potential and search for additional datasets to alleviate the noise of the provided dataset.

In addition, since Document QA is good for prototyping, it was a good idea to use it at the beginning to check that their hypothesis of modeling dataset retrieval as a Question Answering task was right. However, at some point during the project, they should have changed it to another model with state-of-the-art performance. Also, they use symbolic queries for the MRQA model. But since they are generating specific queries for each publication, it should be possible to define and generate queries in the form of embeddings. This could help to improve even more the recall boost provided by the Query Generation Module and at the same time avoid the generation of noise. Furthermore, for research fields, other ranking methods should have been tried like BM25, a ranking function used by search engines whose performance is better than TF-IDF. Finally, for research methods, they should have analyzed more the dataset to use more suitable techniques such as unsupervised NER instead of supervised NER.

Conclusion

The KAIST team proposed to model the dataset retrieval task as a Question Answering task. This is a unique approach in this competition and led to successful results as shown in the analysis. This approach is flexible because it allows the retrieval of new datasets that are not in the training set. In addition, the model does not require to be trained on a dataset discovery task. They achieved good results even though they used a pre-trained model on SQuAD, a dataset for Question Answering using Wikipedia pages.

Their approach to retrieve research fields is simple, fast to compute, and powerful. It can retrieve specific research fields with high precision. On the other hand, the proposed approach to retrieve research methods did not achieve as good results as the other task. The main problem was that they did not tackle the noise problem in the dataset.

What comes next

This work is the very first step of the Coleridge Initiative to build an “Amazon.com” for data users and data producers. The next step is to construct a system that recommends datasets to researchers. The KAIST team has the hypothesis that datasets depend on research fields and vice versa. For example, in the research field *Question Answering*, a subfield of *Natural Language Processing* and *Computer Science*, the most commonly used dataset is SQuAD [@rajpurkar2016squad]. Therefore, according to their hypothesis, two publications using SQuAD are presumably to be in the same field, *Question Answering*. Based on this hypothesis, it would be possible to build hierarchical clusters of publications with the same research field. In this way, a cluster will have publications with the same research field and similar datasets. As an example, the QA cluster will have papers about QA and those papers will use similar datasets like SQuAD and TriviaQA [@joshi2017triviaqa]. With these clusters, the system will be able to recommend datasets to data users. For example, if a publication is in the *Question Answering* field, the proposed system would be able to recommend the authors SQuAD and TriviaQA. Moreover, it would be able to recommend to data producers fields with a lack of datasets.

In addition, there is room for improvement in the models they proposed. For example, since they used a pre-trained model in Document QA, they think they did not exploit the whole potential of this system, so they would like to train the model using a big enough training set of publications.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. 2013–0–00179, Development of Core Technology for Context-aware Deep-Symbolic Hybrid Learning and Construction of Language Resources) and Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (2017M3C4A7065962).

Appendix: Description of the code and documentation

The technical documentation of the code is provided in the GitHub repository of the project https://github.com/HaritzPuerto/RCC/tree/master/project

Chapter 8 - Knowledge Extraction from scholarly publications: The GESIS contribution to the Rich Context Competition

Knowledge Extraction from scholarly publications - The GESIS contribution to the Rich Context Competition

Authors: *Wolfgang Otto, Andrea Zielinski, Behnam Ghavimi, Dimitar Dimitrov, Narges Tavakolpoursaleh, Karam Abdulahhad, Katarina Boland, Stefan Dietze*

Affiliation: *GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany*

Corresponding author: *wolfgang.otto@gesis.org*

1. Introduction

GESIS - the Leibniz Institute for the Social Sciences (GESIS)[1] is the largest European research and infrastructure provider for the social sciences and offers research data, services and infrastructures supporting all stages of the scientific process. The GESIS department *Knowledge Technologies for the Social Sciences (WTS)*[2] is responsible for developing all digital services and research data infrastructures at GESIS and aims at providing integrated access to social sciences data and services. Next to traditional social sciences research data, such as surveys and census data, an emerging focus is to build data infrastructures able to exploit novel forms of social sciences research data, such as large Web crawls and archives.

Research at WTS[3] addresses areas such as Information Retrieval (IR), Information Extraction (IE) & Natural Language Processing (NLP), semantic technologies and human computer interaction and aims at ensuring access and use of social sciences research data along the FAIR principles, for instance, through interlinking of research data, established vocabularies and knowledge graphs and by facilitating semantic search across distinct platforms and datasets. Due to the increasing importance of Web- and W3C standards as well as Web-based research data platforms, in addition to traditional research data portals, findability and interoperability of research data across the Web constitutes one current challenge. In the context of Web-scale reuse of social sciences resources, the extraction of structured data about scholarly entities such as datasets and methods from unstructured and semi-structured text, as found in scientific publications or resource metadata, is crucial in order to be able to uniquely identify social sciences resources and to understand their inherent relations.

Prior works at WTS/GESIS addressing such challenges apply NLP and machine learning techniques to, for instance, extract and disambiguate mentions of datasets[4] (Boland et al., 2012; Ghavimi et al., 2016)), authors (Backes, 2018a, 2018b) or software tools (Boland and Krüger, 2019) from scientific publications or to extract and fuse scholarly data from large-scale Web crawls (Sahoo et al., 2017; Yu et al., 2019). Resulting pipelines and data are used to empower scholarly search engines such as the *GESIS-wide search*[5] (Hienert et al., 2019) which provides federated search for scholarly resources (datasets, publications etc.) across a range of GESIS information systems or the *GESIS DataSearch* platform[6] (Krämer et al., 2018), which enables search across a vast number of social sciences research datasets mined from the Web.

Given the strong overlap of our research and development profile with the recent initiatives of the Coleridge Initiative to evolve this research field through the Rich Context Competition (RCC)[7], we are enthusiastic about having participated in the RCC2018 and are looking forward to continue this collaboration towards providing sound frameworks and tools which automate the process of interlinking and retrieving scientific resources.

The central tasks in the RCC are the extraction and disambiguation of mentions of datasets and research methods as well as the classification of scholarly articles into a discrete set of research fields. After the first phase, each team received feedback from the organizers of the RCC consisting of a quantitative and qualitative evaluation. Whereas quantitative results of our inital contribution throughout phase one have shown significant room for improvement, the qualitative assessement, conducted by four judges on a sample of ten documents, underlined the potential of our approach.

Since we have been shortlisted for the second phase of the RCC, this chapter describes our approaches, techniques, and additional data used to address all three tasks. As described in the following subsections, we decided to follow a module-based approach where each module or the entire pipeline can be reused. The remaining chapter is organised as follows. The following Section 2 provides an overview of our approach, used background data and preprocessing steps, whereas Sections 3, 4, and 5 describe our approaches in more detail, including results towards each of the tasks. Finally, we discuss our results in Section 6 and provide an overview of future work in Section 7.

2. Approach, data and pre-processing

This section describes the external data sources we used as well as our pre-processing steps.

2.1 Approach overview and initial evaluation feedback

The central tasks in the RCC are the extraction of dataset mentions from text. Even so, we considered the discovery of research methods and research fields important. To this end, we decided to follow a module-based approach. Users could choose to use each specific module solely or as parts of a data processing pipeline. Figure 8.2 shows an overview of modules developed and their dependencies. Here, the upper three modules (which are in gray) describe the pre-processing steps (cf. Section 2.3). The lower four modules (blue) are used to generate the output in a predefined format as specified by the competition.

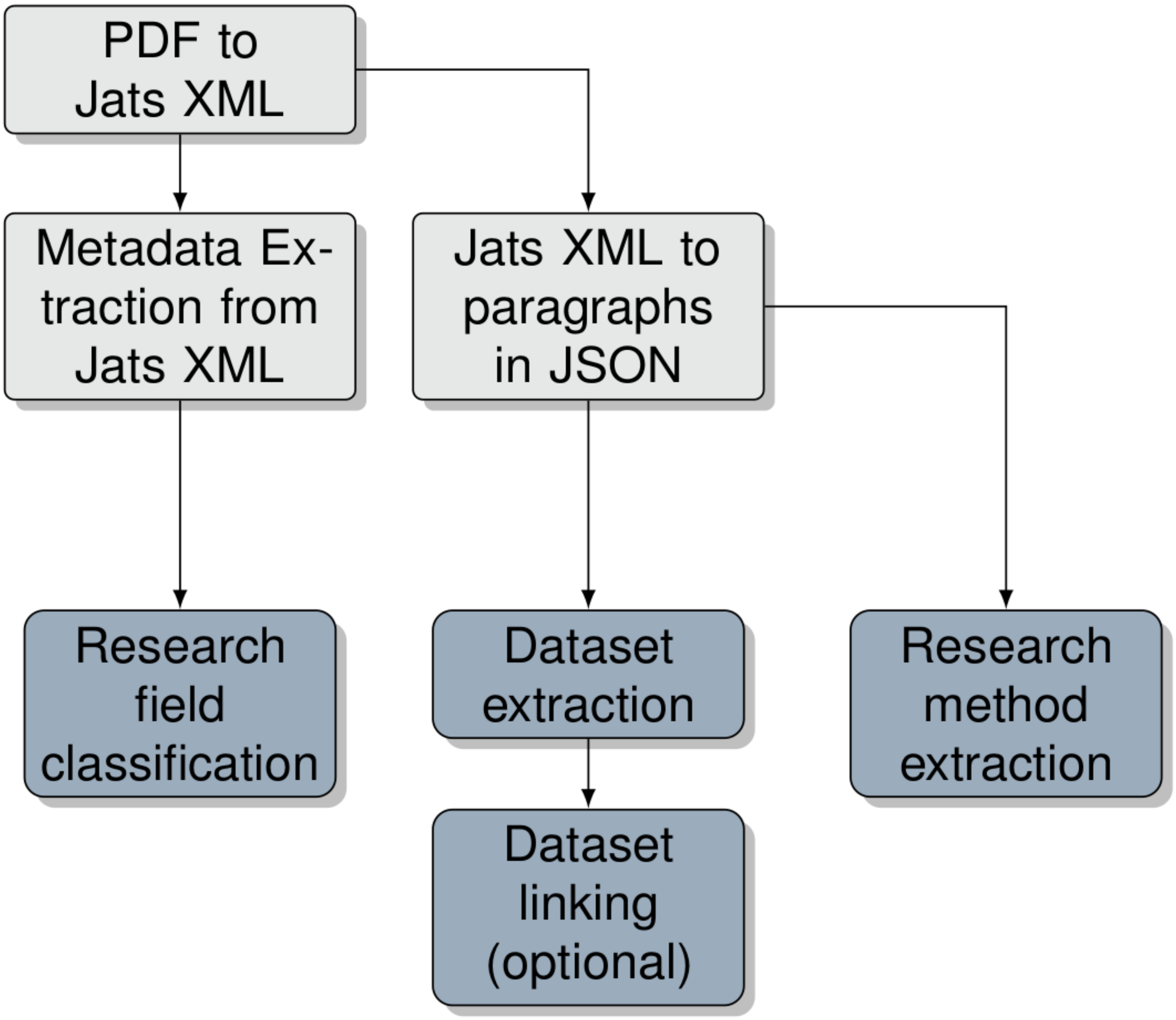

PDF to Jats XML
Metadata Extraction from Jats XML
Jats XML to paragraphs in JSON
Research field classification
Dataset extraction
Research method extraction
Dataset linking (optional)

Figure 8.1: An overview of the individual software modules described in this document and their dependencies. Gray: Our pre-processing pipeline. Blue: three main tasks of the RCC.

The pre-processing step consists of extracting metadata and raw text from PDF documents. The output of this step is then used by the software modules responsible for tackling the individual sub-tasks. These sub-tasks are to discover research datasets (cf. Section 3), methods (cf. Section 4) and fields (cf. Section 5). First, a Named Entity Recognition module is used to find dataset mentions. This module used a supervised approach trained on a weakly labled corpus. In the next step, we combine all recognized mentions for each publication and compare these mentions to the metadata from the list of datasets given by the competition. For this linking step the mentions and year information located in the same sentence are used. The corresponding sentence and extracted information are saved for debugging and potential usage in future pipeline components. The task of identifying research methods is solved unsing a Named Entity Recognition and Linking module with incorporated word embeddings and lexical resources. For identifying research fields, we trained a classifier on openly available abstracts and metadata from the domain of social sciences crawled from the Social Science Open Access Repository[8] (SSOAR). We tried different classifiers and selected the best performing one, a classifier based on fasttext[9], i.e. a neural net based approach with a high performance(Joulin et al., 2017).

After the first phase, each team received feedback from the organizers of the RCC. The feedback is two folds, a quantitative and qualitative evaluation. Unfortunately, the quantitative assessment showed our algorithm for dataset mention retrieval did not perform well regarding precision and recall metrics. In contrast to this, our approach has been found convincing regarding the quality of results. The qualitative feedback is based on a random sample of ten documents given to four judges. The judges were asked to manually extract dataset mentions. After this the overlap between their dataset extractions and the output of our algorithm was calculated. Other factors that judges took into consideration are specificity, uniqueness, and multiple occurrences of dataset mentions. As for the extraction of research methods and fields, no ground truth has been provided; these tasks were evaluated against the judges' expert knowledge. Similarly to the extraction of dataset mentions, specificity and uniqueness have been considered for these two tasks. The feedback our team received was overall positive.

2.2 External data sources

For developing our algorithms, we utilized two external data sources. For the discovery of research methods and fields, we resort to data from the Social Science Open Access Repository[10] (SSOAR). GESIS – Leibniz Institute for the Social Sciences maintains SSOAR by collecting and archiving literature of relevance to the social sciences.

In SSOAR, full texts are indexed using controlled social science vocabulary (Thesaurus[11], Classification[12]) and are assigned rich metadata. SSOAR offers documents in various languages. The corpus of English language publications that can be used for purposes of the competition consists of a total of 13,175 documents. All SSOAR documents can be accessed through the OAI-PMH[13] interface.

Another external source we have used to discover research methods is the ACL Anthology Reference Corpus (Bird et al., 2008). ACL ARC is a corpus of scholarly publications about computational linguistics. The corpus consists of a total of 22,878 articles.

2.3 Pre-processing

Although the organizers of the RCC offered plain texts for the publication, we decided to build our own pre-processing pipeline. The extraction of text from PDF files is still an error prone process. To handle de-hyphenation and paragraph segmentation during extraction time and benefit from automatic metadata extraction (i.e. title, author, abstracts and references) we decided to use a third party extraction tool. The Cermine Extraction Tool[14](Tkaczyk et al., 2015) transforms the files into XML documents using the Journal Article Tag Suite[15](Jats). For the competition we identified two interesting elements of the Jats XML format, i.e., (<)front(>) and (<)body(>). The (<)front(>) element contains the metadata of the publication, whereas the (<)body(>) contains the main textual and graphic content of the publication. As a last step of the pre-processing, we removed all linebreaks from the publication. The output of this step is a list of metadata fields and values, as shown in Table 8.1 for each publication paragraph.

	Example Text Field Data
publication_id	12744
label	paragraph_text
text	A careful reading of text, word
	for word, was …
section_title	Data Analysis
annotations	[{’start’: 270, ’end’: 295,
	’type’: ’bibref’, …
section_nr	[3, 2]
text_field_nr	31
para_in_section	1

Table 8.1: Example preprocessing output for a paragraph in a given publication.

3. Dataset extraction

3.1 Task description

In the scientific literature, datasets are cited to reference, for example, the data on which an analysis is performed or on which a particular result or claim is based. In this competition, we focus on

(i) extracting and (ii) disambiguating dataset mentions from social science publications to a list of given dataset references. Identifying dataset mentions in literature is a challenging problem due to the huge number of styles of citing datasets. Although there are proposed standards for dataset citation in full-texts, researchers still ignore or neglect such standards (see, e.g., (Altman and King, 2007)). Furthermore, in many research publications, a correct citation of datasets is often missing (Boland et al., 2012). The following two sentences exemplify the problem of the usage of an abbreviation to make a reference to an existing dataset. The first example illustrates the use of abbreviations that are known mainly in the author's research domain. The latter illustrates the ambiguity of abbreviations. In this case, *WHO* identifies a dataset published by the World Health Organization and does not refer to the institution itself.
Example 1: *P-values are reported for the one-tail paired t-test on* Allbus* (dataset mention) and *ISSP* (dataset mention).*
Example 2: *We used* WHO data from 2001* (dataset mention) to estimate the spreading degree of AIDS in Uganda.*
We treat the problem of detecting dataset mentions in full-text as a Named Entity Recognition (NER) task.

Formal problem definition

Let (D) denote a set of existing datasets (d) and the knowledge base (K) as a set of known dataset references (k). Furthermore, each element of (K) is referencing an existing dataset (d). The Named Entity Recognition and Linking task is defined as (i) the identification of dataset mentions (m) in a sentence, where (m) references a dataset (d) and (ii) linking them, when possible, to one element in (K) (i.e., the reference dataset list given by the RCC).

3.2 Challenges

We focus on the extraction of dataset mentions in the body of the full-text of scientific publications. There are three types of dataset mentions: (i) The full name of a dataset (”National Health and Nutrition Examination Survey“), (ii) an abbreviation (”NHaNES“) or (iii) a vague reference, e.g., ”the monthly statistic“. With all these these types, the NER task faces special challenges. In the first case, the used dataset name can vary in different publications. For instance one publication cites the dataset with ”National Health and Nutrition Examination Survey“ the other could use the words ”Health and Nutrition Survey“. In the case where abbreviations are used, a disambiguation problem occurs, e.g., in ”WHO data“. WHO may describe the World Health Organization or the White House Office. In the case, that an abbreviation is used after the dataset name has been written in full, the mapping between these different spellings in one text is referred to as Coreference Resolution. The biggest challenge is again the lack of annotated training data. In the following we describe how we have dealt with this lack of ground truth data.

3.3 Phase one approach

Missing ground truth data is the main problem to handle during this competition. To this end, supervised learning methods for dataset mentions extraction from texts are not applicable without the identification of external training data or the creation of useful labeled training data from information given by the competition. Because of the lack of existing training data for the task of dataset mention extraction we resort to the provided list of dataset mentions and publication pairs and re-annotate the particular sentences in the publication text. A list of dataset identifying words is provided for some of the known links between publications and datasets by the competition. These words represent the evidence of the linkage between publication and datasets and are extracted from the publication text. In the course of re-annotation, we search for each of the identifying words in the corresponding publication texts. For each match, we annotate the occurence in our raw text and use these annotations as ground truth. As described in the pre-processing section, our units for processing the publication text are paragraphs. The re-annotated corpus consists of a list of paragraphs for each publication with stand-off annotations identifying the mentions of datasets (i.e. position of the start and end characters and the entity type for each mention: *dataset*). This re-annotation is then used to train Spacy's neural network-based NER model[16]. We created a holdout set of 1,000 publications and a training set of size 4,000. Afterwards, we train our model with the paragraphs as a sampling unit. In the training set, 0.45 percent of the paragraphs contained mentions. For each positive training example, we have added one negative sample that contains no known dataset mentions and is randomly selected. We used a batch size of 25 and a dropout rate of 0.4. The model was trained for 300 iterations.

Evaluation

We evaluated our model with respect to four metrics: precision and recall, each for strict and for partial match. While the strict match metrics are standard evaluation metrics, the partial match metrics are their relaxed variants in which the degree to which dataset mentions have to match can vary. Consider the following partial match example: “National Health and Nutrition Examination Survey” is the extracted dataset mention, while “National Health and Nutrition Examination Survey (NHANES)” is the true dataset mention. In contrast to the strict version of the metrics, this overlapping match is considered a match for the partial version. The scores describe whether a model is able to find the correct positions of dataset mentions in the texts, even if the start and end positions of the characters are not the same, but the ranges overlap.

Metric	Value
Precision (partial match)	0.93
Recall (strict match)	0.95
Precision (strict match)	0.80
Recall (strict match)	0.81

Table 8.2: Performance of phase one approach of dataset extraction.

Table 8.2 shows the results of the dataset mention extraction on the holdout set. The model can achieve high strict precision and recall values. As expected, the results are even better for the partial version of the metrics. It means that even if we couldn’t match the dataset mention in a text exactly, we can find the right context with very high precision.

3.4 Phase two approach

In the second phase of the competition, additional 5,000 publications were provided by RCC. We extended our approach to consider the list with dataset names supplied by the organizers and re-annotated the complete corpus of 15,000 publications in the same manner as in phase one to obtain training data. This time we split the data in 80% for training and 20% for test.

Evaluation

We resort to the same evaluation metrics as in phase one. However, we calculate precision and recall on the full-text of the publication and not on the paragraphs as in the first phase.
Table 8.3 shows the results achieved by our model. We observe lower precision and recall values. Compared to phase one, there is also a smaller difference between the precision and recall values for the strict and partial version of the metrics.

Metric	Value
Precision (partial match)	0.51
Recall (partial match)	0.90
Precision (strict match)	0.49
Recall (strict match)	0.87

Table 8.3: Performance of phase two approach for dataset extraction.

4. Research method extraction

4.1 Task description

Inspired by a recent work of Nasar et al. (Nasar et al., 2018), we define a list of basic entity types that give key-insights into scholarly publications. We adapted the list of semantic entity types to the domain of the social sciences with a focus on *research methods*, but also including related entity types such as *theory, todel, measurement, tool, performance*. We suspect that the division into semantic types might be helpful to find *research methods*. The reason is that the related semantic entities types might provide clues or might be directly related to the research method itself.

For example, in order to achieve a certain research goal, an experiment is used in which a certain combination of *methods* is applied to a *dataset*. The methods can be specified as concepts or indirectly through the use of certain *software*. The result is then quantified with a *performance* using a specific measure.

Example: *P-values* (measurement) are reported for the *one-tail paired t-test* (method) on *Allbus* (dataset) and *ISSP* (dataset).
We selected the entity types *research method*, *research theory*, *research tool* and *research measurement* as the target research method

related entity types (see Table 8.4). This decision is based
on an ecxamination of the SAGE ontology given by the RCC as a sample of how research method terms might look like.

Entity type	Corresponding SAGE type	Examples
Research Method	SAGE-METHOD	Bootstrapping, Active Interviews
Research Measurement	SAGE-MEASURE	Latent Variables, Phi coefficient, Z-score
Research Theory	SAGE-THEORY	Frankfurt school, Feminism, Actor network theory
Research Tool	SAGE-TOOL	SPSS, R statistical package

Table 8.4: Entity types of relevance for the research method extraction task.

Formal problem definition

The task of Named Entity Recognition and Linking is to (i) identify the mentions (m) of research-related entities in a sentence and (ii) link them, if possible, to a reference knowledge base (K) (i.e. the SAGE Thesaurus[17]) or (iii) assign a type to each entity, e.g. a *research method*, selected from a set of predefined types.

4.2 Challenges

There are some major challenges that any named entity recognition, classification and linking system needs to handle. First, regarding NER, identifying the entities boundary is important, thus detecting the exact sequence span. Second, ambiguity errors might arise in classification. For instance,'range' might be a domain-specific term from the knowledge base or belong to the general domain vocabulary. This is a challenging task for which context information is required. In the literature, this relates to the problem of **domain adaptation** which includes fine-tuning to specific named entity classes[18]. With respect to entity linking, another challenge is detecting name variations, since entities can be referred to in many different ways. Semantically similar words, synonyms or related words, which might be lexically or syntactically different, are often not listed in the knowledge base (e.g., the lack of certain terms like 'questioning' but not 'questionnaire'). This problem of automatically detecting these relationships is generally known as **linking problem**. Note that part of this problem also results from PDF-to-text conversion which is error-prone. Dealing with incomplete knowledge bases, i.e. **handling of out of vocabulary (OOV) items**, is also a major issue, since knowledge bases are often not exhaustive enough and do not cover specific terms or novel concepts from recent research. Last but not least, the combination of different semantic types gives a more coherent picture of a research article. We hypothesize that such information would be helpful and results in an insightful co-occurrence statistics, and provides additional detail directly related to entity resolution, and finally helps to assess the **relevance of terms** by means of a score.

4.3 Our approach

Our research method extraction tool builds on Stanford's CoreNLP and Named Entity Recognition System[19]. The information extraction process follows the workflow depicted in figure 8.2, using separate modules for pre-processing, classification, linking and term filtering.

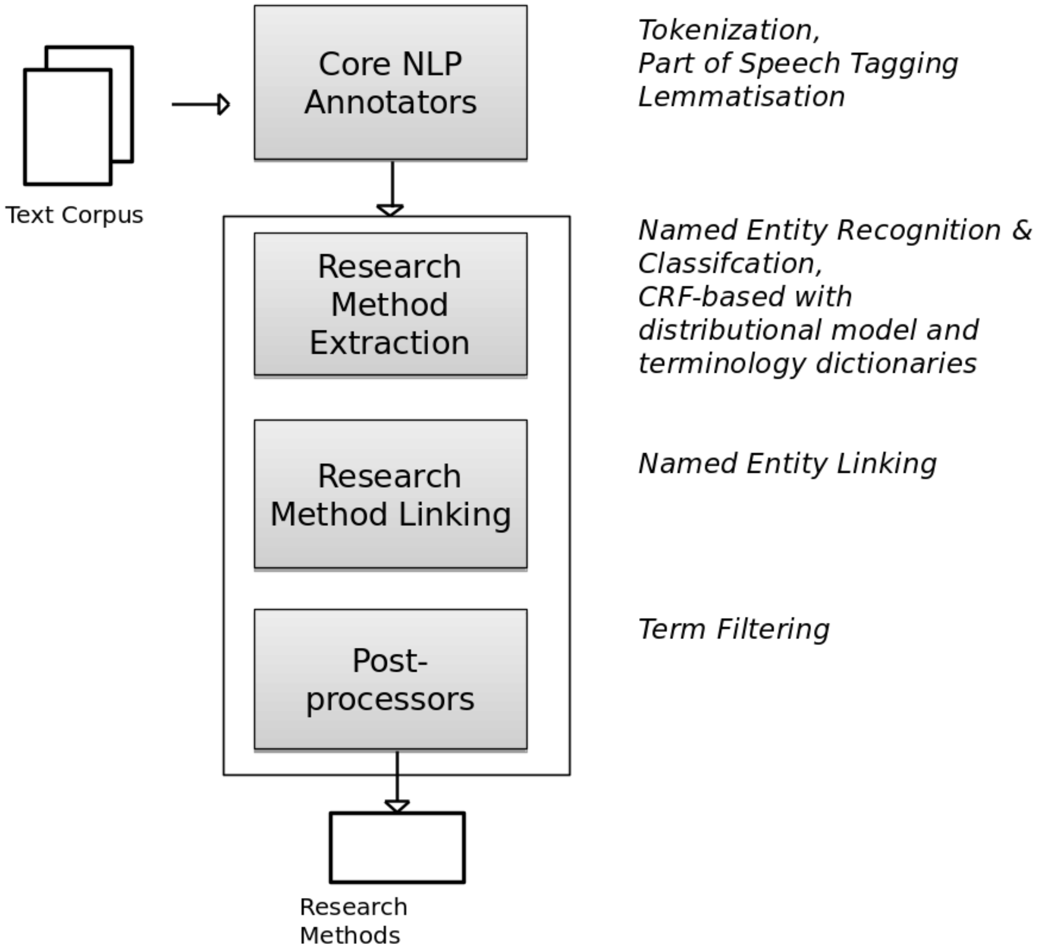


Figure 8.2: Overview of the entity extraction pipeline.

We envision the task of finding entities in scientific publications as a sequence labeling problem, where each input word is classified as being of a dedicated semantic type or not. In order to handle entities related to our domain, we train a CRF based machine learning classifier with major semantic classes, (see Table 8.4, using training material from the ACL RD-TEC 2.0 dataset (QasemiZadeh and Schumann, 2016). Apart from this, we follow a domain adaptation approach inspired by (Agerri and Rigau, 2016) and ingest semantic background knowledge extracted from external scientific corpora, in particular the ACL Anthology (Bird et al., 2008; Gildea et al., 2018). We perform entity linking by means of a new gazetteer based on th SAGE dictionary of Social Research Methods (Lewis-Beck et al., 2003), thus putting a special emphasis on the social sciences. The linking component addresses the synonymy problem and matches an entity despite name variations such as spelling variations. Finally, term filtering is carried out based on termhood and unithood, while scoring is achieved by calculating a relevance score based on TF-IDF (cf Table 8.6).

Our research experiments are based on publications from the Social Science Open Access Repository (SSOAR)[20] as well as the train and test data of the Rich Context Competition corpus[21]. Our work extends previous work on this topic (cf. (Eckle-Kohler et al., 2013)) in various ways: First, we do not limit our study to abstracts, but use the entire fulltext. Second, we focus on a broader range of semantic classes, i.e. *Research Method*, *Research Theory*, *Research Tool* and *Research Measurement*, tackling also the problem of identifying novel entities.

Distributed semantic models

For domain adaptation, we integrate further background knowledge. We use topical information from word embeddings trained on an scientific corpus as an additional feature to our NER model. For this, we use agglomerative clustering of the word embeddings to identify topical groups of words. The cluster number of each word is used as additional sequential input feature for our CRF model. Semantic representations of words are a successful extension of common features, resulting in higher NER performance (Turian et al., 2010) and can be trained offline. In this work, the word vectors were learned based on 22,878 documents of the scientific ACL Anthology Reference Corpus[22] using Gensim[23] with the skip-gram model (cf. (Mikolov et al., 2013)) and a pre-clustering algorithm[24].

Features

The features incorporated into the linear chain CRF are shown in the Table 8.5. The features depend mainly on the observations and on pairs of adjacent labels, using a log-linear combination. However, since simple token level training of CRFs leads to poor performance, more effective text features such as word shape, orthographic, gazetteer, Part-Of-Speech (POS) tags, along with word clustering have been used.

Type	Features
Token unigrams	(w{i–2}), (w{i–1}), (w{i}), (w{i+1}), (w{i+2}), … | | ***POS unigrams*** | (p{i}), (p{i–1}), (p{i–2})
Shapes	shape and capitalization
NE-Tag	(t{i–1}), (t{i–2})
WordPair	((p{i}), (w{i}), (c{i})) | | ***WordTag*** | ((w{i}), (c_{i}))
Gazetteer	SAGE Gazetteer
Distributional Model	ACL Anthology model

Table 8.5: Features used for NER.

Knowledge resources

We use the SAGE thesaurus which includes well-defined concepts, an explicit taxonomic hierarchy between concepts as well as labels that specify synonyms of the same concept. A portion of terms is unique to the social science domain (e. g., ‘dependent interviewing’), while others are drawn from related disciplines such as statistics (e. g., ‘conditional likelihood ratio test’)[25]. However, since the thesaurus is not exhaustive and covers only the top-level concepts related to social science methods, our aim was to extend it by automatically extracting further terms from domain-specific texts, in particular from the Social Science Open Access Repository. More concretely, we carried out the following steps to extend SAGE as an off-line step. For step 2 and 3, candidate terms have been extracted by our pipeline for the entire SSOAR corpus.

1. Assignment of semantic types to concepts (manual)
2. Extracting terms variants such as abbreviations, synonyms, related terms from SSOAR (semi-automatic)
3. Computation of term and document frequency scores for SSOAR (automatic)

Extracting term variants such as abbreviations, synonyms, and related terms

26,082 candidate terms have been recognized and classified by our pipeline and manually inspected to a) find synonyms and related words that could be linked to SAGE, and b) build a post-filter for incorrectly classified terms. Moreover, abbreviations have been extracted using the algorithm of Schwartz and Hearst (Schwartz and Hearst, 2003). This way, a Named Entity gazetteer could be built and is used at run-time. It comprises 1,111 terms from SAGE and 447 terms from the used glossary of statistical terms[26] as well as 54 previously unseen terms detected by the model-based classifier.

Computation of term and document frequency scores

Term frequency statistics have been calculated off-line for the entire SSOAR corpus. The term frequency at corpus level will be used at run time to determine the term relevance at the document level by calculating the TF-IDF scores. The most relevant terms from SAGE are listed in Table 8.6.

SAGE Term	TF-IDF Score	Semantic Class
Fuzzy logic	591,29	Research Method
arts-based research	547,21	Research Method
cognitive interviewing	521,13	Research Method
QCA	463,13	Research Method
oral history	399,68	Research Method
market research	345,37	Research Field
life events	186,61	Research Field
Realism	314,34	Research Theory
Marxism	206,77	Research Theory
ATLAS.ti	544,51	Research Tool
GIS	486,01	Research Tool
SPSS	136,52	Research Tool

Table 8.6: Most relevant terms from SAGE by Semantic Type.

Definition of a relevance score

Relevance of terminology is often assessed using the notion of *unithood*, i.e. ‘the degree of strength or stability of syntagmatic combinations of collections’, and *termhood*, i.e. ‘the degree that a linguistic unit is related to domain-specific concepts’ (Kageura and Umino, 1996). Regarding *unithood*, the NER model implicitly contains heuristics about legal POS tag sequences for candidate terms, consisting of at least one noun (NN), preceeded or followed by modifiers such as adjectives (JJ), participles (VB*) or cardinal numbers (CD), complemented by wordshape features.

In order to find out if the candidate term also fulfills the *termhood* requirement, domain-specific term frequency statistics have been computed on the SSOAR repository, and set in contrast to general domain vocabulary terms. It has to be noted that only a small portion of the social science terms is actually unique to the domain (e.g., ‘dependent interviewing’), while others might be drawn from related disciplines such as statistics (e.g., ‘conditional likelihood ratio test’).

Preliminary results

Our method has been tested on 100 fulltext papers from SSOAR and ten documents from the Rich Context Competition (RCC), all randomly selected from a hold out corpus. In our experiments on SSOAR Social Science publications, we compared results to the given metadata information. The main finding was that while most entities from the SAGE thesaurus could be extracted and linked reliably (e.g., ’Paired t-test’), they could not be easily mapped to the SSOAR metadata terms, which consist of only a few abstract classes (e.g., ’quantitative analysis’). Furthermore, our tool was tested by the RCC organizer, were the judges reviewed ten random publications and generated qualitative scores for each document. In this evaluation, the research method extraction tool received the overall best results of all competitors for this task.[27]

5. Research field classification

5.1 Task description

The goal of this task is to identify the research fields covered in the social science publications. In general, two approaches could be applied to this task. One is the extraction of relevant terms of the publications. It means that this task could be seen as a keyword extraction task and the detected terms considered as descriptive terms regarding the research field. The second approach is to learn to classify publications research fields with the use of annotated data in a superviesed manner. The benifit of the second approach is that the classification scheme to describe the research field can be defined by experts of the domain. The disadvantage of supervised trained classifiers for this task is the lack of applicable training data. Furthermore, it must be ensured that the training data is comparable to the texts the research field classifier should be applied on.

Formal problem definition

Let (P) denote a set of publications of size (n), (A) a set of corresponding abstracts of the same size and (L) a set of (k) defined class labels describing research fields. The task of research field classification is to select for each publication (p_i\in{P}) based on the information contained in the corresponding abstract (a_i\in{A}) a set of labels (C_i = \varnothing \cap {c_1\dots c_n|c_n \in L}) of (n) labels. The number of (n) denotes the number of labels from (L) describing the research field of (a_i) and can vary for each publication (p_i). If there is no label (l_k) representing the information given by the abstract (a_i) the set of class labels is the empty set (\varnothing).

5.2 Our approach

Since we didn’t receive any gold standard for this task during the competition we decided to make use of external resources. We decided to use an external labeled dataset to train a text classifier which is able to predict one or moreresearch label for a given abstract of a publication.

The publications given througout the competition belongs to the domain of social sciences we considered metadata from a open access repository for the social sciences called SSOAR. The advantages are twofold. On the one hand, we could rely on professional annotations in a given classification scheme covering the social sciences and related areas. On the other hand the source is openly available.[28]

The annotated data of SSOAR contains four different annotation schemes for research field related information. By reviewing these schemes, we decided to use the Classification Social Science (classoz) annotation scheme. The number of classes in each schema, coverage of each classification, and the distribution of data in each schema affected our decision. An exhausitve description of the used data can be found in Section 8.2.

Pre-processing and model architecture

SSOAR is a multilingual repository. Therefore, the available abstracts may vary in language and the language of the abstract may differ from the language of the article itself. We selected all English abstract with valid classification as our dataset. Mainly because of the language of the RCC corpus. However, it should be noted that the multilingual SSOAR abstract corpus has a skewed distribution of languages with English and German as the main languages. We count 22,453 English abstracts with valid classification after filtering. Due to the unequal distribution of labels in the dataset, we need to guaranty enough training data for each label. We selected only labels with frequency over 300 for training the model, which results in a total of 44 out of 154 classification labels representing research fields. For creating train and test set, 22,453 SSOAR publications with their assigned labels were split randomly. We used a train/validation/test split of 70/10/20. We decided to train a text classifier based on a fasttext (Joulin et al., 2017) model in the author's implementation. The arguments to use this model was the speed in comparison to a more complex neural net architecture and the still comparable to state of the art performance (e.g.(Wang et al., 2018)). The model is trained with learning rate 1.0 for 150 epochs. Also, the negative sampling parameter is set to 25.

5.3 Evaluation

Figure 8.3 shows the performance of the model regarding various evaluation metrics for different thresholds. A label is assigned to a publication if the model outputs a probability for the label above the defined threshold. In multi-label classification, this allows us to evaluate our model from different perspectives. As illustrated in figure 8.3, the intersection of the micro precision and the micro recall curves is at the threshold of 0.1, where the highest micro f1 score is achieved. By increasing the threshold from this point, the micro-precision score is increasing, but the micro recall is falling. By decreasing threshold, these trends are inverted. Also, the default threshold of 0.5 doesn't look promising. In spite of micro-precision about 0.75, we have a problem with the very high number of items without any prediction. In respect to this observation it is advantageous to select a lower threshold in a productive environment. The curve named *without prediction* shows for a given threshold the share of publications in the test set without any prediction. If the selected threshold value is high, the number of publications for which the model cannot predict a research field increases. For example, a selected threshold value of 0.55 leads to 40% unclassified publications in the test set. The *one correct* named curve indicates the quality of the publication wise prediction. It shows the share of all publications in the test set where at least one of the predicted research field labels can be found in the ground truth data. For instance, if a threshold of 0.1 is selected for 75% of the publications in the test set, at least one of the model predictions are correct. This value decreases with increasing threshold simmilar to the recall metric. The final micro f1 value on the test set for our model and a selected threshold of 0.1 is 0.56 (precison 0.55, recall 0.56).

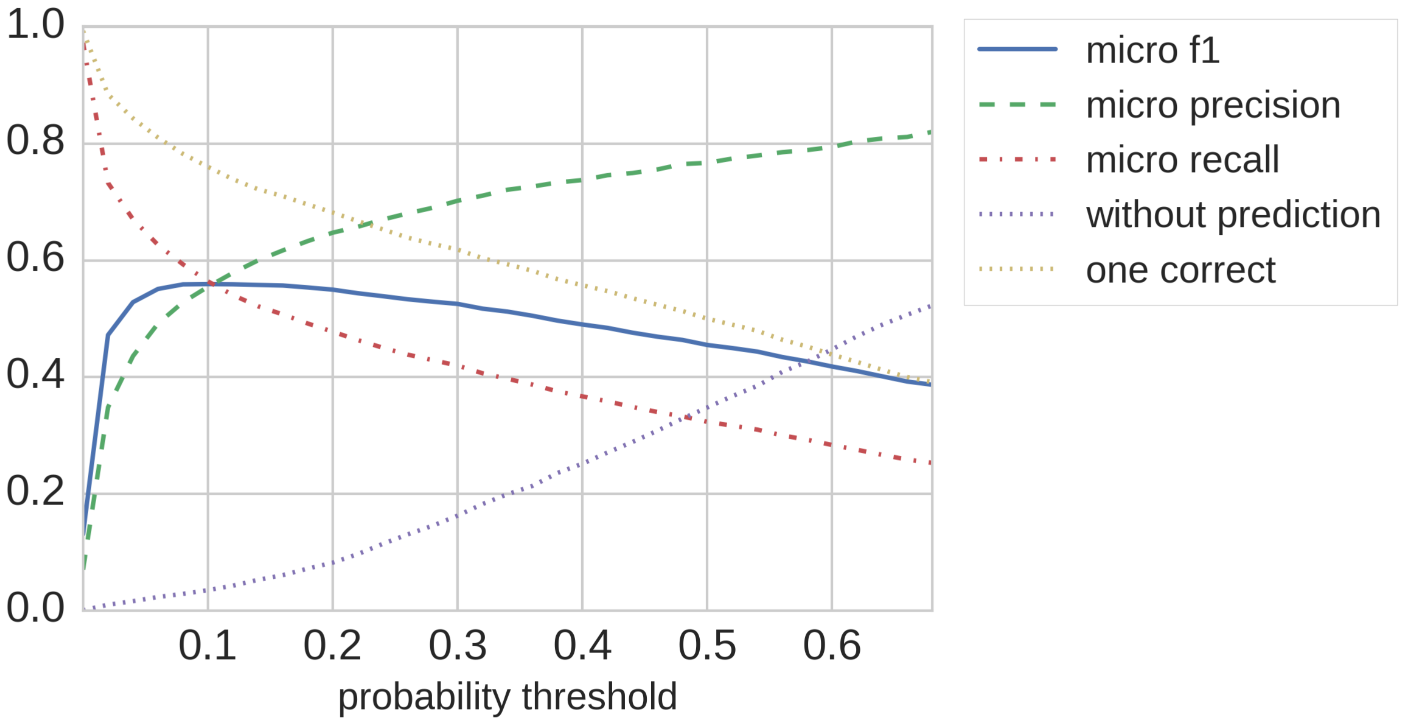


Figure 8.3: Performance for different selected probability thresholds (validation set).

6. Discussions and Limitations

6.1 Dataset Extraction.

For the dataset extraction task, the proposed methods are only tested on social science related data. The performance measures we have introduced are based on a hold out data set of our automatically created dataset. Especially the recall may be biased given that our training as well as testing data is biased towards known datasets, where datasets not yet part of our reference set are not considered.

The results of the second phase presented during the RCC workshop[29] are showing good performance of our approach in comparison to the other finalist teams with the highest precision 52.2% (second: 47.0%) and second in recall (ours: 20.5, best: 34.8%). With respect to F1, our approach provides the second best performing system for this task (29.5%, 40.0% for first place). The results on the manually created hold out set underline, that our system performs better in respect to precision in comparison to the other finalist teams. Given that our models are supervised through a corpus of social sciences publications, we anticipate limited generalisability across other disciplines and plan to investigate this aspect as part of future work. In this context, the focus of our training data towards survey data, also reflected in dataset titles such as *Current Population Survey*, could have biased the model to detect the survey as a specific type of research datasets better than other subtypes like e.g. text corpora in the NLP community. In general, however, our approach to using a weakly labeled corpus created from a list of dataset names could be applied in other research domains.

6.2 Research method extraction.

We consider the extraction of research methods from full text as a particularly challenging task because the sample vocabulary given by the RCC organizers covers a large thematic variety of areas. The task itself was defined as the identification of research methods associated with a specific publication, which in turn are drawn from a specific research field. Since no training data has been provided, we created and annotated a new corpus for the task and trained a CRF model, adding lexical resources. The qualitative reviews during the two phases of the competition attested that this approach works fine.

6.3 Research field classification.

Our supervised machine learning approach to handle the research field classification task performs well on the dataset created from social science publication metadata. A micro F1 measure of above 55% seems to indicate reasonable performance considering the small dataset with 44 labels and a mean number of keywords of three terms per publication. As one example of multilabel classification with a comparable size of labels we would like to mention the classification of texts in the domain of medicine presented in (Wang et al., 2018). The models tested by the authors on the task of multilabel prediction from 50 different labels leads to micro F1 values between 53% and 62%.
Considering the evaulation approach, focused on publications from the social sciences, the generalisability across other disciplines remains unclear and requires further research. Even though the used classification scheme may cover neighbouring disciplines, for instance, medicine, the numbers of samples of the training data covering other research fields than the social science is limited. Our pragmatic approach of basing our classifications on the abstract of the publications makes it applicable even in scenarios where the full-text of publications is not accessible.

7. Conclusion

This chapter has provided an overview on our solutions submitted to the Rich Context Competition 2018. Aimed at improving search, discovery and interpretability of scholarly resources, we are addressing three distinct tasks all aimed at extracting structured information about research resources from scientitifc publications, namely the extraction of dataset mentions, the extraction of mentions of research methods and the classification of research fields.

In order to address all aforementioned challenges, our pipelines make use of a range of preprocessing techniques together with state-of-the-art NLP methods as well as supervised machine learning approaches tailored towards the specific nature of scholarly publications as well as the dedicated tasks. In addition, background datasets have been used to facilitate supervision of methods at larger scale.

Our results indicate both significant opportunities for automating the aforementioned three tasks but also their challenging nature, in particular given the lack of publicly available gold standards for training and testing. Aggregating and publishing such data has been identified as important activity for future work, and is a prerequisite for significantly advancing state-of-the-art methods.

Acknowledgments

This work has been partially funded by Deutsche Forschungsgemeinschaft (DFG) under grant number MA 3964/7–1.
Wolfgang Otto acknowledges the enabling support provided by the Indo-German Joint Research Project titled ‘Design of a Sciento-text Computational Framework for Retrieval and Contextual Recommendations of High-Quality Scholarly Articles’ (Grant No. DST/INT/FRG/DAAD/P–28/2017) for this work.

References

Agerri R and Rigau G (2016) Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence* 238. Elsevier: 63–82.

Altman M and King G (2007) A proposed standard for the scholarly citation of quantitative data. *D-lib Magazine* 13(3/4).

Backes T (2018a) Effective unsupervised author disambiguation with relative frequencies. In: *JCDL* (eds J Chen, MA Gonçalves, JM Allen, et al.), 2018, pp. 203–212. ACM. Available at: http://dblp.uni-trier.de/db/conf/jcdl/jcdl2018.html#Backes18.

Backes T (2018b) The impact of name-matching and blocking on author disambiguation. In: *CIKM* (eds A Cuzzocrea, J Allan, NW Paton, et al.), 2018, pp. 803–812. ACM. Available at: http://dblp.uni-trier.de/db/conf/cikm/cikm2018.html#Backes18.

Bird S, Dale R, Dorr BJ, et al. (2008) The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In: *Proceedings of the sixth international conference on language resources and evaluation (lrec 2008)*, 2008. European Language Resources Association (ELRA).

Boland K and Krüger F (2019) Distant supervision for silver label generation of software mentions in social scientific publications. In: *Proceedings of the 4th joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries*, 2019, pp. 15–27.

Boland K, Ritze D, Eckert K, et al. (2012) Identifying references to datasets in publications. In: *International conference on theory and practice of digital libraries*, 2012, pp. 150–161. Springer.

Eckle-Kohler J, Nghiem T-D and Gurevych I (2013) Automatically assigning research methods to journal articles in the domain of social sciences. In: *Proceedings of the 76th asis\&T annual meeting: Beyond the cloud: Rethinking information boundaries*, 2013, p. 44. American Society for Information Science.

Finkel JR, Grenager T and Manning C (2005) Incorporating non-local information into information extraction systems by gibbs sampling. In: *Proceedings of the 43rd annual meeting on association for computational linguistics*, 2005, pp. 363–370. Association for Computational Linguistics.

Ghavimi B, Mayr P, Lange C, et al. (2016) A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use* 36(3–4). IOS Press: 171–187.

Gildea D, Kan M-Y, Madnani N, et al. (2018) The acl anthology: Current state and future directions. In: *Proceedings of workshop for nlp open source software (nlp-oss)*, 2018, pp. 23–28.

Hienert D, Kern D, Boland K, et al. (2019) A digital library for research data and related information in the social sciences. In: *JCDL* (eds M Bonn, D Wu, JS Downie, et al.), 2019, pp. 148–157. IEEE. Available at: http://dblp.uni-trier.de/db/conf/jcdl/jcdl2019.html#HienertKBZM19.

Joulin A, Grave E, Bojanowski P, et al. (2017) Bag of tricks for efficient text classification. In: *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 2, short papers*, 2017, pp. 427–431. Association for Computational Linguistics.

Kageura K and Umino B (1996) Methods of automatic term recognition: A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3(2). John Benjamins Publishing Company: 259–289.

Krämer T, Klas C-P and Hausstein B (2018) A data discovery index for the social sciences. In: *Scientific data*, 2018.

Lewis-Beck M, Bryman AE and Liao TF (2003) *The Sage Encyclopedia of Social Science Research Methods*. Sage Publications.

Mikolov T, Sutskever I, Chen K, et al. (2013) Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*, 2013, pp. 3111–3119.

Nasar Z, Jaffry SW and Malik MK (2018) Information extraction from scientific articles: A survey. *Scientometrics* 117(3). Springer: 1931–1990.

QasemiZadeh B and Schumann A-K (2016) The acl rd-tec 2.0: A language resource for evaluating term extraction and entity recognition methods. In: *LREC*, 2016.

Sahoo P, Gadiraju U, Yu R, et al. (2017) Analysing structured scholarly data embedded in web pages. *Lecture Notes in Computer Science* 9792. Springer.

Schwartz AS and Hearst MA (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. In: *Pacific symposium on biocomputing*, 2003, pp. 451–462.

Tkaczyk D, Szostek P, Fedoryszak M, et al. (2015) CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)* 18(4). Springer: 317–335.

Turian J, Ratinov L and Bengio Y (2010) Word representations: A simple and general method for semi-supervised learning. In: *Proceedings of the 48th annual meeting of the association for computational linguistics*, Stroudsburg, PA, USA, 2010, pp. 384–394. ACL '10. Association for Computational Linguistics. Available at: http://dl.acm.org/citation.cfm?id=1858681.1858721.

Wang G, Li C, Wang W, et al. (2018) Joint embedding of words and labels for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. DOI:10.18653/v1/p18–1216.

Yu R, Gadiraju U, Fetahu B, et al. (2019) KnowMore - knowledge base augmentation with structured web markup. *Semantic Web* 10(1): 159–180. Available at: http://dblp.uni-trier.de/db/journals/semweb/semweb10.html#YuGFLRD19.

1. https://www.gesis.org/en/institute
2. https://www.gesis.org/en/institute/departments/knowledge-technologies-for-the-social-sciences/
3. https://www.gesis.org/en/research/applied-computer-science/labs/wts-research-labs
4. https://www.gesis.org/en/research/external-funding-projects/archive/infolis-i-and-ii
5. https://search.gesis.org
6. https://datasearch.gesis.org/
7. https://coleridgeinitiative.org/richcontextcompetition
8. https://www.ssoar.info
9. https://fasttext.cc/
10. https://www.gesis.org/ssoar/home
11. https://www.gesis.org/en/services/research/tools/thesaurus-for-the-social-sciences
12. https://www.gesis.org/angebot/recherchieren/tools-zur-recherche/klassifikation-sozialwissenschaften
    (in German)
13. http://www.openarchives.org
14. https://github.com/CeON/CERMINE
15. https://jats.nlm.nih.gov
16. https://spacy.io
17. http://methods.sagepub.com
18. apart from those used in traditional NER systems like *Person*, *Location*, or *Organization* with abundant training data, as covered in the Stanford NER system(Finkel et al., 2005)
19. https://nlp.stanford.edu/projects/project-ner.shtml

20. https://www.ssoar.info
21. https://coleridgeinitiative.org/richcontextcompetition with a total of 5,000 English documents
22. https://acl-arc.comp.nus.edu.sg/
23. https://radimrehurek.com/gensim/
24. Word embeddings are trained with a skip gram model using embedding size equal to 100, word window equal to 5, minimal occurrences of a word to be considered 10. Word embeddings are clustered using agglomerative clustering with a number of clusters set to 500, 600, 700. Ward linkage with Euclidean distance is used to minimize the variance within the clusters.
25. A glossary of statistical terms as provided in https://www.statistics.com/resources/glossary/ has been added as well.
26. Based on https://www.statistics.com/resources/glossary
27. Rank: 1,2,2,1,1 for judges 1–5.
28. A script to download the metadata of SSOAR can be found in github/research-field-classifier
29. Agenda of the Workshop: https://coleridgeinitiative.org/richcontextcompetition/workshopagenda. The results of the finalists are presented here: https://youtu.be/PE3nFrEkwoU?t=9865.

Chapter 9 - Finding datasets in publications: The University of Paderborn approach

Abstract

The steadily increasing number of publications available to researchers makes it difficult to keep track of the state of the art. In particular, tracking the datasets used, topics addressed, experiments performed and results achieved by peers becomes increasingly tedious. Current academic search engines render a limited number of entries pertaining to this information. However, having this knowledge would be beneficial for researchers to become acquainted with all results and baselines relevant to the problems they aim to address. With our participation in the NYU Coleridge Initiative's Rich Context Competition, we aimed to provide approaches to automate the discovery of datasets, research fields and methods used in publications in the domain of Social Sciences. We 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, we used word embeddings. In this chapter, we describe how our approaches performed, their limitations, some of the encountered challenges and our future agenda.

```
author:
- Rricha Jalota
- Nikit Srivastava
- Daniel Vollmers
- René Speck
- Michael Röder
- Ricardo Usbeck
- ‘Axel-Cyrille [Ngonga Ngomo]{}’
bibliography:
- ‘references.bib’
title:
‘Finding datasets in publications: The University of Paderborn approach’
```

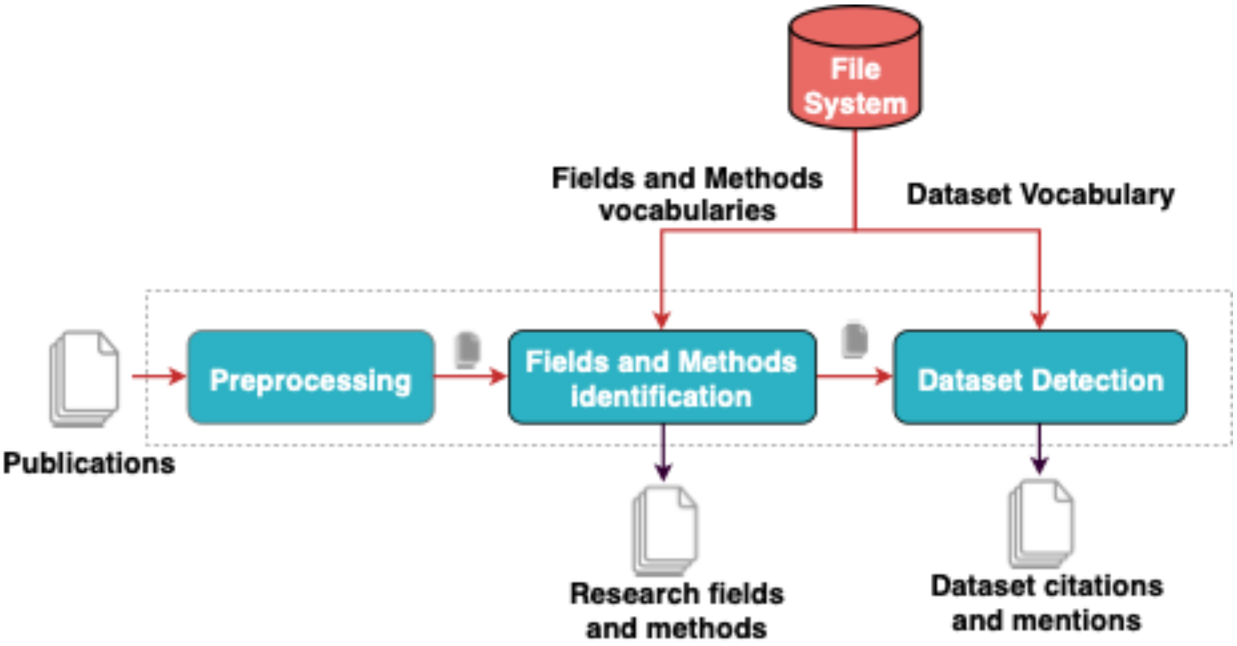
Literature Review

Previous works on information retrieval from scientific articles are mainly seen in the field of Bio-medical Sciences and Computer Science, with systems [@DBLP:journals/ploscb/WestergaardSTJB18] built using the MEDLINE[1] abstracts, full-text articles from PubMed Central[13] or ACL Anthology dataset[14]. The documents belonging to the above-mentioned datasets follow a similar format, and thus, several metadata and bibliographical extraction frameworks like CERMINE [@tkaczyk2014cermine] have been built on them. However, since articles belonging to the domain of Social Sciences do not follow a standard format, extracting key sections and metadata using already existing frameworks like GROBID [@lopez2009grobid], ScienceParse[15] or ParsCit [@councill2008parscit] did not seem as viable options, majorly because these systems were still under development and lacked certain desired features. Hence, building upon the approach of Westergaard et. al [@DBLP:journals/ploscb/WestergaardSTJB18], we built our own sections-extraction framework for dataset detection and research fields and methods identification.

Apart from content and metadata extraction, key-phrase or topic extraction from scientific articles has been another emerging research problem in the domain of information retrieval from scientific articles. Jansen et al. [@jansen2016extracting] extracted core claims from scientific articles by first detecting keywords and key-phrases using rule-based, statistical, machine learning and domain-specific approaches and then applying document summarization techniques. For characterizing a research work in terms of its focus, application domain and techniques used, Gupta et al. [@gupta2011analyzing] proposed applying semantic extraction patterns to the dependency trees of sentences in an article's abstract. On the other hand, to thematically represent scientific articles and for ranking the extracted key-phrases, Mahata et al. [@mahata2018key2vec] devised an approach for processing text documents to train phrase embeddings.

The problem of dataset detection and methods and fields identification is not only different from the ones mentioned above, but also our approach for tackling it is radically disparate. The following sections describe our approach in detail.

Project Architecture


File System
Fields and Methods vocabularies
Dataset Vocabulary
Preprocessing
Fields and Methods identification
Dataset Detection
Publications
Research fields and methods
Dataset citations and mentions

{width="\textwidth"}

Our pipeline (shown in Figure [fig:flowchart]) consisted of three main components: 1) Preprocessing, 2) Fields and Methods Identification and 3) Dataset Extraction. The Preprocessing module read the text from publications and generated some additional files (see Section [preprocess] for details). These files along with the given Fields and Methods vocabularies were used to infer Research Fields and Methods from the publications. Then, the information regarding Research Fields was passed onto the Dataset Detection module and using the Dataset Vocabulary, Dataset Citations and Mentions were identified. The following sections provide a detailed overview of each of these components.

Preprocessing

As discussed in Chapter 5, the publications were provided in two formats: PDF and text. For Phase–1, we used the given text files, however during Phase–2, we came across many articles in the training files that had not been properly converted to text and contained mostly non-ASCII characters. To work with such articles, we relied on the open source tool `pdf2text` from `poppler suite`[16] to extract text from PDFs. The `pdf2text` command served as the first preprocessing step and was called as a subprocess from within a python script. It was used with `-nopgbrk` argument to generate the text files.

Once we had the text files, we followed the rule-based approach as proposed by Westergaard et al. [@DBLP:journals/ploscb/WestergaardSTJB18] for pre-processing. The following series of operations based mostly on regular expressions were performed:

- Words split by hyphens were de-hyphenated
- Irrelevant data was removed (i.e., equations, tables, acknowledgment, references);
- Main sections (i.e., abstract, keywords, JEL-Classification, methodology/data, summary, conclusion) were identified and extracted;
- Noun phrases from these sections were extracted (using the python library, spaCy[17]).

We came up with the heuristics for identifying the main sections after going through the articles from different domains in the training data. We collected the surface forms for the headings of all major sections (abstract, keywords, introduction, data, approach, summary, discussion) and applied regular expressions to search for them and separate them from one another. The headings and their corresponding content were stored as key-value pairs in a file. For generating noun-phrases, this file was parsed and for all the values (content) in key-value (heading-content) pairs, a spaCy object, `doc`, was created sentence-wise. Using the built-in function for extracting noun chunks ([`doc.noun_chunks`]{}), we generated key-value pairs of heading and noun-phrases found in the content and stored them in another file. This file was later used for fields and methods identification.

To determine how well our approach performed in distinguishing sections, we evaluated it on the articles in the validation dataset. During evaluation, we figured out the limiting cases of our approach. A section could not be differentiated either when there was no explicit mention of any of its surface forms or if there were multiple mentions of the surface forms in the articles. For instance, in the validation dataset (see Table [tab:sections]), keywords were not extracted from 13 articles because of no explicit mention of the term ’keywords’ or its variants. On manual inspection, we found keywords were actually not mentioned in these 13 articles. In the remaining articles where the keywords were present, our algorithm could not detect them from 1 article. For brevity, we have reported only four main sections in Table [tab:sections]: title, abstract, keywords and methodology/data, since these are the ones getting preferential treatment in methods and fields identification. If a section was not found in the article (because of no explicit mention of any of the surface forms), then only the sections that could be detected were extracted. The remaining content was saved as `reduced_content` after cleaning and noun-phrases were extracted from it to prevent loss of any meaningful data.

[C[4cm]{} C[3.5cm]{} C[4cm]{}]{} **Sections** & **No explicit mention** & **Mentioned but not distinguished**
Title & 0 & 0
Keywords & 13 & 1
Abstract & 0 & 1
Methodology/Data & 18 & 4

In addition to the main sections, we also extracted PDF metadata using `pdfinfo` service from the `poppler suite` library. The metadata very often contained the keywords and subject of an article, which was helpful in those cases where the keywords were not found by the regular expression.
In the end, the preprocessing module generated four text files for a publication: PDF-converted text, PDF-metadata, processed articles containing relevant data, and noun phrases from the relevant sections, respectively. These files were then passed on to the other two components of the pipeline, which have been discussed below.

Approach

Research Fields and Methods Identification

Vocabulary Generation and Model Preperation

1. **Research Methods Vocabulary**: In Phase–1 of the challenge, we used the given methods vocabulary. However, the feedback that we received from Phase–1 evaluation gave more emphasis to statistical methods used by the authors, references to the time scope, unit of observation, and regression equations rather than the means used to compile the data, i.e., surveys. Since the given methods vocabulary was not a complete representation of statistical methods and also consisted terms depicting surveys, in Phase–2, we decided to create our own Research Methods Vocabulary using Wikipedia and DBpedia.[18] We manually curated a list of all the relevant statistical methods from Wikipedia[19] and fetched their descriptions from the corresponding DBpedia resources. For each label in the vocabulary, we extracted noun phrases from its description and added them to the vocabulary. Please refer Table [tab:vocab] for examples.

   [C[1.5cm]{} C[5cm]{} C[5cm]{}]{} **Label** & **Description** & **Noun Phrases from Description**
   Political forecasting & Political forecasting aims at predicting the

outcome of elections. & Political forecasting, the outcome, elections
Nested sampling algorithm & The nested sampling algorithm is a computational approach to the problem of comparing models in Bayesian statistics, developed in 2004 by physicist John Skilling. & algorithm, a computational approach, the problem, comparing models, Bayesian statistics, physicist John Skilling

2. **Research Fields Vocabulary**: For both the phases, we used the given research fields vocabulary and, just like the methods vocabulary, supplemented it with the noun phrases from the description of the research field labels. However, since our phase–1 model seemed to confuse fields with methods, for Phase–2, we additionally created a stopword-list of terms that didn't contain any domain-specific information, such as; Mixed Methods, Meta Analysis, Narrative Analysis and the like.
3. **Word2Vec Model generation**: In this pre-processing step, we used the above-mentioned vocabulary files containing noun phrases to generate a vector model for both research fields and methods. The vector model was generated by using the labels and noun phrases from the description of the available research fields and methods to form a sum vector. The sum vector was basically the sum of all the vectors of the words present in a particular noun phrase. 3em [The pre-trained Word2Vec model `GoogleNews-vectors-negative300.bin` [@DBLP:journals/corr/abs–1301–3781] was used to extract the vectors of the individual words.]{}
4. **Research Method training results creation**: For research methods, we generated an intermediate result file with the publications present in the training data. It was generated using a `naïve finder algorithm` which, for each publication, selected the research method with the highest cosine similarity to any of its noun phrase's vectors. This file was later used to assign weights to research methods using Inverse Document Frequency.

Processing with Trained Models

- **Finding Research Fields and Methods:** To find the research fields and methods for a given list of publications, we performed the following steps: (At first, Step 1 was executed for all the publications, thereafter Step 2 and 3 were executed iteratively for each publication).
    1. **Naïve Research Method Finder run** - In this step, we executed the `naïve research method finding algorithm` (i.e. selected a research method based on the highest cosine similarity between vectors) against all the current publications and then merged the results with the existing result from the `research methods’ preprocessing step`. The combined result was then used to generate IDF weight values for each `research method`, to compute the significance of recurring terms.
    2. **IDF-based Research Method Selection** - We re-ran the algorithm to find the closest research method to each noun phrase and then sorted the pairs based on their weighted cosine similarity. The weights were taken from the IDF values generated in the first step and the manual weights assigned (section-wise weighting). Here, the noun phrases that came from the methodology section and from the methods listed in JEL-classification (if present) were given a higher preference. The pair with the highest weighted cosine similarity was then chosen as the Research Method of the article.
    3. **Research Field Finder run** - In this step, we first found the closest research field from each noun phrase in the publication. Then we selected the Top N (= 10) pairs that had the highest weighted cosine similarity. Afterwards, the noun phrases that had a similarity score less than a given threshold (= 0.9) were filtered out. The end-result was then passed on to a post-processing algorithm.

       For weighted cosine similarity, the weights were assigned manually based on the section of publication from which the noun phrases came. In general, noun phrases from title and keywords (if present) were given a higher preference than other sections, since usually these two sections hold the crux of an article. Note, if sections could not be discerned from an article, then noun phrases from the section, reduced_content (see section [preprocess]), were used to find both fields and methods.
    4. **Research Field Selection** - The top-ranked term from the result of step 3, which was not present in the stopword-list of irrelevant terms, was marked as the research field of the article.

The experimental set-up and average training times (ATT) have been reported in Table [tab:setup]:

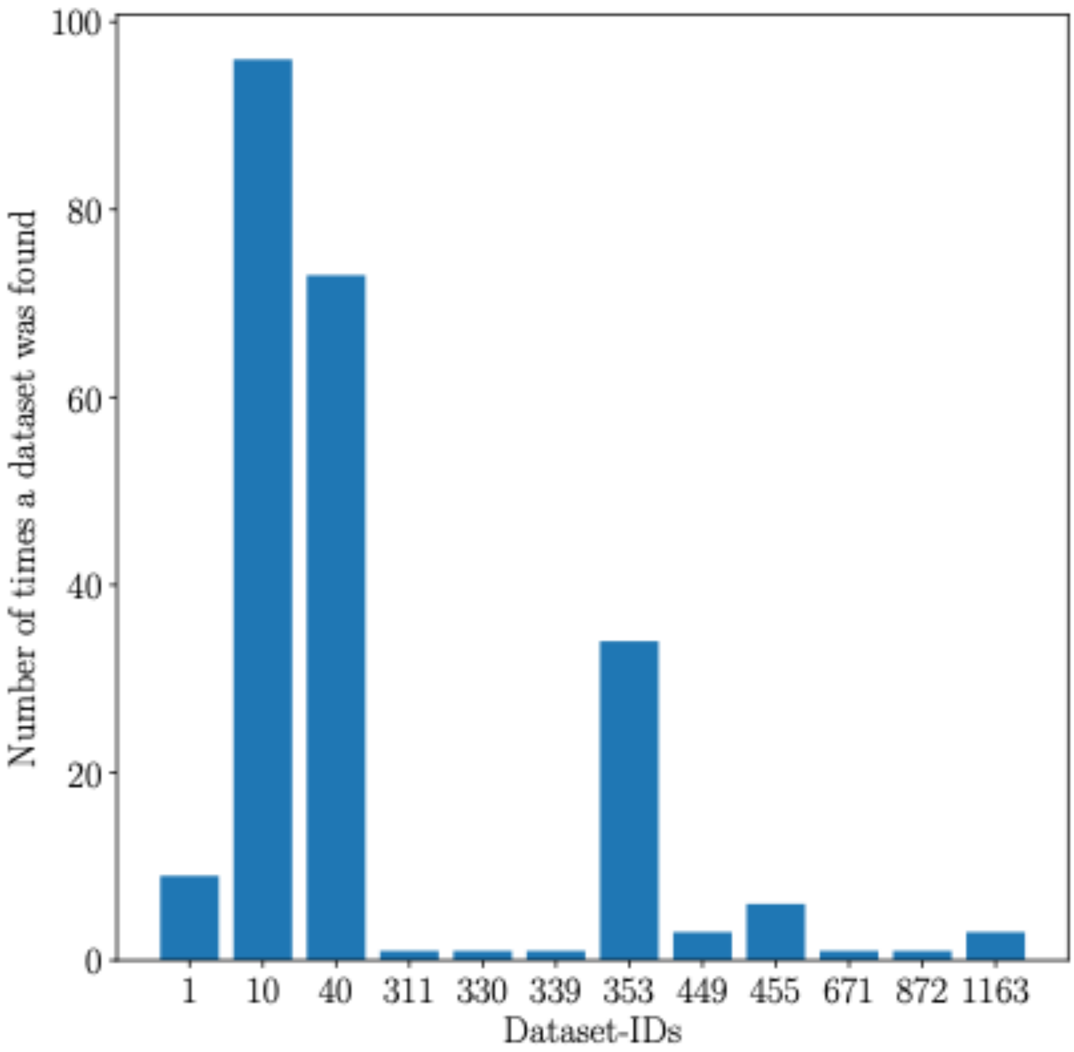
[|C[5cm]{} | C[7cm]{} |]{} Computing Infrastructure & macOS, 2 GHz Intel Core i7 processor, 4 cores RAM 16 GB 1600 MHz
ATT - RF model & 3m 21s
ATT - RM model & 3m 19s
Link to Implemented Code &
https://github.com/nikit91/Jword2vec/tree/rich-context

Dataset Extraction

For identifying the datasets in a publication, we followed two approaches and later combined results from both. Both the approaches have been described below.

1. **Simple Dataset Mention Search:** We chose the dataset citations from the given Dataset Vocabulary that occurred for one dataset only and used these unique mentions to search for the corresponding datasets using regular expressions in the text documents. Then, we computed a frequency distribution of the datasets. As can be seen from Figure [fig:graph], certain dataset citations occurred more often than others. This is because while searching for dataset citations, apart from the dataset title, the corresponding mention_list from Dataset Vocabulary was also considered, which contained many commonly occurring terms like ‘time’, ‘series’, ‘time series’, ‘population’ etc. Therefore, we filtered out those dataset citations that occurred more than a certain threshold value (=1.20) multiplied by the median of the frequency distribution and that had less than 3 distinct mentions in a publication. The remaining citations were written to an interim result file.
Table [tab:simple] depicts the improvement in performance of Simple Dataset Mention Search with the inclusion of filtering. The filtering process improved the F1-measure by 42.86%. Note, as the validation data consisted of only 100 articles, changing the threshold value to 1.10 or 1.30 didn’t result in any significant change, hence we have maintained a constant threshold value of 1.20 in our comparison table.

   [C[1.5cm]{} C[3.5cm]{} C[3.5cm]{} C[3.5cm]{}]{} **Metrics** &
   without filtering & **Threshold=1.20, mentions $<$ 3** &
   Threshold=1.20, mentions $<$ 4
   Precision & 0.09 & 0.71 & 0.09
   Recall & 0.28 & 0.12 & 0.28
   F1-score & 0.14 & **0.20** & 0.14



Frequency Distribution of Dataset Citations[]{data-label="fig:graph"}

2. **Rasa-based Dataset Detection:** In our second approach, we trained an entity extraction model based on conditional random fields (CRF) using Rasa NLU [@DBLP:journals/corr/abs–1712–05181]. For training the model we used the Spacy Tokenizer[20] for the preprocessing step. For Entity Recognition we used BILOU tagging and used 50 iterations to train the CRF. We used the Part of Speech tags, the case of the input tokens and the suffixes of the tokens as input features for the CRF model. We particularly tested two configurations for training the CRF-based Named Entity Recognition (NER) model. In Phase–1, the 2500 labeled publications from the training dataset were used for training the Rasa NLU[21] model. Later in Phase–2, when the Phase–1 holdout corpus was released, we combined its 5000 labeled publications with the previously given 2500 labeled publications and then retrained the model again with these 7500 labeled publications.
   Running the CRF-Model: The trained model was run against the preprocessed data to detect dataset citations and mentions. Only the entities that had a confidence score greater than a certain

threshold value (= 0.72) were considered as dataset mentions. A dataset mention was considered as a citation only if it was found in the given Dataset Vocabulary (via string matching either with a dataset title or any of the terms in a dataset ‘mention_list’) and if it belonged to the research field of the article. To check if a dataset belonged to the field of research, we found the cosine similarity of the terms in the ‘subjects’ field of the Dataset Vocabulary with the keywords and the identified Research Field of the article.

3. **Combining the two approaches:** The output generated by the Rasa-based approach was first checked for irrelevant citations before a union was performed to combine the results. This was done by checking if a given dataset_id occured more than a threshold value (= 1.20) multiplied by the median of the frequency distribution (same as the filtering process of the Simple Dataset Mention Search).

Note that, the threshold values mentioned above were set after some experiments of trial and testing. For dataset extraction, the goal was to keep the number of false positives low while not compromising the true positives. For research methods and fields, a manual evaluation (see the next section for details) was done to test if the results made sense with the articles.

Evaluation

We performed a quantitative evaluation for Dataset Extraction using the evaluation script provided by the competition organizers (refer Chapter 5 for more details). This evaluation (see Table [tab:dataset]) was carried out against the validation data, wherein we compared four different configurations. As can be inferred from the table, there was only a slight increase in performance for the Rasa-based model, when the training samples were increased. However, combining it with the Simple Dataset Mention Search, increased the performance by *19.42%*. Interestingly, there was no improvement in performance in the combined approach even when the training samples for the Rasa-based model were increased. This might be because of the removal of frequently-occuring terms from the Rasa-generated output, based on the frequency distribution of dataset mentions as computed in the Simple Dataset Mention Search.

[M[2.2cm]{} | M[2.3cm]{} | M[2.2cm]{} M[2.2cm]{} M[2.2cm]{}]{} & & **Metrics**& **Rasa-based Approach** (2500) & **Rasa-based Approach** (7500) & **Combined Approach** (2500) & **Combined Approach** (7500)

Precision & 0.382 & 0.388 & **0.456** & **0.456**
Recall & 0.26 & 0.26 & **0.31** & **0.31**
F1 & 0.309 & 0.311 & **0.369** & **0.369**

For Research Fields and Methods, we carried out a qualitative evaluation against 10 randomly selected articles from Phase–1 holdout corpus. Tables [tab:field] and [tab:method] depict a comparison between the predicted fields and methods in Phase–1 and Phase–2. In general, our models returned a more granular output in the second phase, solely because of the modifications we made in the vocabularies.

[C[1cm]{} C[4.5cm]{} C[3cm]{} C[3.5cm]{}]{} **pubid** & **Keywords** & **Phase–1** & **Phase–2**
10328 & Cycling for transport, leisure and sport cyclists & Health evaluation & **Public health and health promotion**
7270 & Older adult drug users, harm reduction & Health Education & **Correctional health care**
6053 & Economic conditions - crime relationship, homicide & Homicide & **Gangs and crime**

[C[1cm]{} C[4.5cm]{} C[3cm]{} C[3.5cm]{}]{} **pubid** & **Keywords** & **Phase–1** & **Phase–2**
10328 & Thematic content analysis & Thematic analysis & **Sidak correction**
7270 & Interviews conducted face to face, finding systematic patterns or relationships among categories identified by reading the interview transcript & Qualitative interviewing & **Sampling design**
6053 & Autoregressive integrated moving average (ARIMA) time-series model & Methodological pluralism & **Multivariate statistics**

Discussion

Throughout the course of this competition, we encountered several challenges and limitations in all the three stages of the pipeline. In the preprocessing step, the appropriate extraction of text from PDFs turned out to be rather challenging. This was especially due to the varied formats of the publications, which made the extraction of specific sections—that contained all data relevant to our work—demanding. As mentioned before, if there was no explicit mention of the key-terms like `Abstract, Keywords, Introduction, Methodology/Data, Summary, Conclusion` in the text, then the content was saved as 'reduced_content' after applying all other preprocessing steps and filtering out any irrelevant data.
Our experiments suggest that the labeled publications we received for dataset detection were not uniform in the dataset mentions provided, which made it difficult to train an entity extraction model even with an increased number of training samples. Hence, there was only a slight improvement in performance when the Rasa-model was trained with 7500 publications instead of 2500. This was also why we combined the Rasa-based approach with the Simple Dataset Mention Search, so that at least the datasets that were present in the vocabulary do not get missed.

Regarding the fields and methods, vocabularies played an immense role in their identification. The vocabularies that were provided by the SAGE publications contained some terms that were either polysemous or very high-level and therefore, were picked up by our model very often. Hence, for research methods, we created our own vocabulary containing all the relevant statistical methods, and for fields, we introduced a stopword-list of irrelevant terms and looked it up each time, before writing the result to the output file. The goal of stopword-list generation was to filter the terms that did not carry domain-specific information and sounded more like research methods than fields. Since the focus was on more granulated results, we tried to look for open ontologies for Social Science Fields and Methods and unfortunately, could not find any. It is worth mentioning that since our approach for Fields and Methods identification relied heavily upon vocabularies, it could not find any new methods or fields from the publications.

Based on the final evaluation feedback, since our Phase–2 models did not perform as good as we expected, following are a few things that we could have done differently.

1. For research methods, merging the given SAGE methods vocabulary with our manually curated vocabulary, could have resulted in methods that would have been both granular and statistical while still being relevant to the publications. Introducing a stopword-list just as we did for research field identification, could also have been another workaround.
2. For both fields and methods identification, we could have also tried pre-trained embeddings from glove[22] and fastText[23].
3. As our entity-extraction approach for Dataset Detection suffered from a limitation of inconsistent labels (i.e. datasets mentioned in the form of abbreviation, full-name, collection procedure, and keywords) in training data, we could have extended the Simple Dataset Mention Search to a pattern-oriented search based on handcrafted rules derived from sentence structure and other heuristics.

Future Agenda

The data provided to us in the competition displayed a cornucopia of inconsistencies even after human processing. We hence propose that machine-aided methods for computing correct and complete structured representation of publications are of central importance for scientific research such as an Open Research Knowledge Graph [@DBLP:journals/corr/abs–1901–10816][@DBLP:conf/esws/BuscaldiDMOR19]. Previous works on never-ending learning have shown how humans and extraction algorithms can work together to achieve high-precision and high-recall knowledge extraction from unstructured sources. In our future work, we hence aim to populate a **scientific knowledge graph** based on never-ending learning. The methodology we plan to develop will be domain-independent and rely on active learning to classify, extract, link and publish scientific research artifacts extracted from open-access papers. Inconsistency will be remedied by ontology-based checks learned from other publications such as SHACL constraints which can be manually or automatically added.[24] The resulting graphs will

- rely on advanced distributed storage for RDF to scale to the large number of publications available;
- be self-feeding, i.e., crawl the web for potentially relevant content and make this content available for processing;
- be self-repairing, i.e., be able to update previous extraction results based on insights gathered from new content;
- be weakly supervised by humans (e.g. authors of publications), who would assist in correcting wrong hypotheses, thereby leveraging semi-supervised learning;
- provide standardized access via W3C Standards such as SPARQL.

Having such knowledge graphs would make it easier for the researchers (both young and veteran) to easily follow along with their domain of fast-paced research and eliminate the need to manually update the domain-specific ontologies for fields, methods and other metadata as new research innovations come up.

Appendix

The code and documentation for all our submissions can be found here: https://github.com/dice-group/rich-context-competition.

Chapter 10 - Finding datasets in publications: The Singapore Management University approach

Finding datasets in publications: The Singapore Management University approach

Simple Extraction for Social Science Publications

Philips Kokoh Prasetyo[1], Amila Silva[2^], Ee-Peng Lim[1], Palakorn Achananuparp[1]
[1]Living Analytics Research Centre, Singapore Management University
[2]University of Melbourne
[^]work done while working at Living Analytics Research Centre
[1]{pprasetyo,eplim,palakorna}@smu.edu.sg,
[2]amila.silva@student.unimelb.edu.au

First draft: 11 February 2019; Second draft: 12 June 2019, Final draft: 29 November 2019

Abstract

With the vast number of datasets and literature collections available for research today, it is very difficult to keep track on the use of datasets and literature articles for scientific research and discovery. Many datasets and research work using them are left undiscovered and under-utilized due to the lack of available search tools to automatically find out who worked with the data, on what research topics, using what research methods and generating what results. The Coleridge Rich Context Competition (RCC) therefore aims to build automated dataset discovery tools for analyzing and searching social science research publications. In this chapter, we describe our approach to solving the first phase of Coleridge Rich Context Competition.

Introduction

Automated discovery from scientific research publications is an important task for analysts, researchers, and learners as they develop the scientific knowledge and use them to gain new insights. More specifically, on the tasks of discovering datasets and methods mentioned in a research publication, we have seen a lack of available tools to easily find who else worked on a particular dataset, what research methods people apply on the dataset, and what results they have found using the dataset. Furthermore, new datasets are not easy to discover, and as a result, good datasets and methods are often neglected.

The Coleridge Rich Context Competition (RCC) aims to build automated datasets discovery from social science research publications, filling the gap of this problem. In this competition, given a corpus of social science research publications, we have to automatically identify datasets used, and then infer the research methods and research fields in the publications. Note that no labeled data are given for research methods and fields identification.

We describes our submission to the first phase of RCC. We perform dataset detection followed by implicit entity linking approach to tackle dataset extraction task. We adopt weakly supervised classification for research methods and fields identification tasks utilizing external resource SAGE Knowledge as proxy for weak labels.

Related Work

Extracting information from scientific text has been explored in the past [PM04; NCKL15; SBP+16]. One type of information extraction from scientific articles is extracting keyphrases and relation between them [ADR+17]. Luan et al. (2017) propose semi-supervised sequence tagging approach to extract keyphrases [LOH17]. Augenstein and Søgaard (2017) explore multi-task deep recurrent neural network approach with several auxiliary tasks to extract keyphrases [AS17].

Another type of extraction is citation extraction. Two citation extraction settings have been explored before: reference mining inside the full text [ACK18], and citation metadata extraction [Het08; APBM14; AGJ+17]. Nasar er al. (2018) write a survey on information extraction from scientific articles [NJM18].

Recently, there are some work to explore dataset extraction from scientific text [BREM12; GML+16; GMVL16]. Boland et al. (2012) propose weakly supervised pattern induction to identify references in social science publications [BREM12]. Ghavimi et al. (2016) propose a semi automatic approach for detecting dataset references for social science texts [GML+16; GMVL16]. Dataset extraction is a challenging task because of the inconsistency and wide range of dataset mention styles in research publications [GMVL16].

Data Analysis

The first phase of RCC dataset consists of a labeled corpus of 5,000 publications for training set, and additional 100 publications for development set. The RCC organizer keeps a separate corpus of 5,000 publications for evaluation. Each article in the dataset contains full text article and dataset citation labels. The metadata of cited datasets in the corpus are also provided. For research methods and fields, no label information is provided, only SAGE social science research method graph and research fields vocabulary are provided. More details on RCC dataset and competition design can be read on chapter 5.

Preprocessing

In order to reliably access important structures of paper publications, we parse all papers using AllenAI Science Parse (https://github.com/allenai/science-parse) [AGB+18]. AllenAI Science Parse reads PDF file, and returns title, authors, abstract, sections, and bibliography (references). Since this parser utilizes machine learning models to parse PDF file, the parsing results may not be 100% accurate. Furthermore, this parser is unable to parse scan copy of old publication. In the situation where we are unable to access parsed fields, we fall back to the given text files.

Mention Analysis

There are 5,499 and 123 dataset citations in training and development set respectively. Among these citations, 320 citations in training set and 6 citations in development set do not have mentions information. We analyze the paper sections where the dataset mentions commonly occur. Table 10.1 and 10.2 show top 12 most common sections mentioning dataset in training and development set. The tables suggest that abstract, reference titles, discussion, results, and methods are the most common sections where the dataset mentions occur. We exploit reference titles for dataset extraction.

Table 10.1: Top 12 Sections Mentioning Datasets in Training Set

Section Header	Mention Frequency
Abstract	2,548
Reference Titles	1,997
Discussion	1,390
Results	836
Methods	804
Introduction	530
Statistical Analysis	285
Comment	279
Acknowledgements	261
Materials and Methods	254
Study Population	227
Data	214

Table 10.2: Top 12 Sections Mentioning Datasets in Development Set

Section Header	Mention Frequency
Abstract	78
Reference Titles	37
Discussion	19
Introduction	14
Results	12
Statistical Analyses	9
Methods	8
Ethics	7
Population	7
Population Impact	7
Price	7
2.1 Data	5

Citation Analysis

We build citation network from training set. Each node in the network is a paper publication, and an edge between two node A and B is generated if a paper A cites paper B. Table 10.3 shows the statistics of the citation network.

Table 10.3: Statistics of Citation Network

Number of nodes	5,000
Number of edges	1,222
Network density	0.0098%

Initially, we propose an approach utilizing citation network based on an intuition that datasets, research methods, and research fields are shared by: 1) same or similar issues, 2) same or similar context, 3) same or similar authors and communities, 4) same or similar metrics used in the publication. However, based on table 10.3, we learn that exploring rich context using paper-paper citation network is not viable at this stage because most papers listed in publications' bibliography are not available in the training set, and therefore, paper-paper citation network becomes very sparse with many unknown information. Figure 10.1 shows the visualization of citation network. As we can see in the citation network visualization, most papers only have one edge, although the average number of papers in bibliography list is 34.5. Furthermore, only 1,466 out of 5,000 publications are not isolated (having at least one bibliography paper in training data). Due to this reason, we drop our idea on utilizing paper-paper citation network at this stage. Nevertheless, we believe that bibliography contains important signals and information about datasets, and research fields.

Figure 10.1: Citation Network from Training Data. Green nodes are publications with dataset citations, and red nodes are publication without dataset citations. Isolated nodes are not visualized.

Methods

In this section, we describe our approach for RCC tasks: dataset extraction, research methods identification, and research fields identification.

Dataset Extraction

We employ a pipeline of two subtasks for dataset extraction: dataset detection, followed by dataset recognition. The goal of dataset detection is to detect whether a publication cites a dataset or not. This first subtask helps us to quickly filter out non-dataset publications. After the first subtask, we mine dataset mentions for the remaining publications in dataset recognition subtask.

For dataset detection, we utilize paper title in bibliography (reference list) combined with explicit research methods mentions to detect whether a publication citing a dataset or not. Explicit research methods mentions are determined based on exact match between paper title and SAGE research methods vocabulary. We train an SVM classifier using explicit research method mentions and n-gram features from paper titles in bibliography. We use the SVM classifier to classify each publication, if the classifier gives positive label, then we proceed to dataset recognition subtask, otherwise we ignore the publication.

For dataset recognition, we use an implicit entity linking approach. We start with the Naive Bayes model, which can be regarded as a standard information retrieval baseline, and entity indicative weighting strategy is used to improve the model. In order to calculate the word distribution of each dataset, we represent each dataset using its title, dataset mentions (provided in the training set), and dataset relevant sentences, filtered from the relevant publications using the rule based approach proposed in [GMVL16]. All these texts related to a particular dataset are considered as a single string, and we calculate the word distribution as follows. Let $\mathbf{w}$ be the set of words in a dataset. In our problem setting, we assume the dataset prior probability $p(d)$ to be uniform. The probability of dataset d given $w \in \mathbf{w}$ is:

$$\begin{aligned} p(d|\mathbf{w}) &\propto \prod_{w \in \mathbf{w}} p(w|d) \\ &= \prod_{w \in \mathbf{w}} \frac{f(d,w)+\gamma}{\sum_{w'} f(d,w')+|W|\gamma} \end{aligned}$$

where $f(d,w)$ is the number of co-occurrences of word w with entity d, γ is the smoothing parameter, and $|W|$ is the vocabulary size. For each dataset d, we derive $f(d,w)$ by the count of w occurrences in the text extracted for each dataset. In order to stress more priority for dataset indicative words, we improved the final objective function of our model as follows:

$$ln(p(d|\mathbf{w})) \propto \sum_{w \in \mathbf{w}} \beta(w) * ln(p(w|d))$$

where $\beta(w)$ is the entity-indicative weight for word w. This weight $\beta(w)$ is added as an exponent to the term $p(w|d)$. $\beta(w)$ is calculated as:

$$\beta(w) = log(1 + E/df(w))$$

where E is the number of distinct datasets considered and $df(w)$ counts the number of datasets with at least one occurrence of w. This model can be trained efficiently, and training the model on RCC dataset needs approximately 5 minutes.

Then for a given unseen publication, we use same rule based approach [GMVL16] to filter a few relevant sentences, and datasets are ranked by $ln(p(d|w))$ to select the most suitable datasets. In order to select exact datasets related to particular publication, we select top 10 datasets ranked using above approach. And then the confidence probability related to the top 10 datasets are normalized and select the datasets with the normalized probability higher than a predefined threshold value. We return the entity indicative words as relevant dataset mentions.

Research Methods Identification

Since we do not have labeled training data for this task, we use explicit research method mentions (based on exact match with SAGE research methods vocabulary) in a publication as weak signals on research methods used in the publication. When these mentions frequently appear in a publication, there is a high chance that this publication is using these particular research methods.

Based on this intuition, we generate training set for research method classification utilizing sentences that explicitly mention research method in a publication. Publication title and the sentences mentioning research method serve as context information of a specific research method. In order to reduce noisy weak signals, we apply minimum support of three sentences in a publication. We exclude research methods which only being mentioned one or two times in a publication. We also exclude research methods that only being mentioned in less than 10 different publications from the training set. Finally, we have 133 research methods having sufficient context information for training data. This number is 20.18% of 659 research methods in SAGE research method graph.

We use the training data to train logistic regression classifier to classify research methods from publication title and sentences. We utilize n-gram features from publication title and sentences for the classifier. We apply the logistic regression classifier to recommend top 3 research methods based on logistic regression probability score.

This approach can be extended by utilizing research method graph to expand the context. Context information does not only comes from sentences in publication, but also comes from related research methods as well as broader concept information. By using this information, we can potentially expand to more than 133 research methods and perform more accurate prediction.

Research Fields Identification

Similar to research methods identification, this task does not have labeled training data. We only have access to list of SAGE research fields. SAGE research fields are organized hierarchically into three levels, namely L1, L2, and L3, for example: Soc–2–4 (*kinship*) is under Soc (*sociology*) in L1, and under Soc–2 (*anthropology*) in L2.

To gain more understanding about the characteristic of each field, we crawl top search results from SAGE Knowledge[2]. From the search result snippets, we collect information such as title and abstract on various publications including case, major work, books, handbooks, and dictionary. We exclude video and encyclopedia. Due to sparseness of the SAGE Knowledge, we exclude all research fields with less than 10 search results. In the end, we have samples of 414 L3 research fields under 101 L2 research fields and 10 L1 research fields. This numbers cover 20.87% of 1,984 L3 research fields, and 67.79% of 149 L2 research fields in the list of SAGE research fields. We use this data to train research fields classifiers.

We build three SVM classifiers for L1, L2, and L3 to classify a publication using paper title and abstract. Instead of taking the highest score, we take top-k research fields and perform re-ranking considering agreement among L1, L2, L3. We return a research field if its upper level are also in top ranks. Since level L1 is too general, we only output research fields from L2, and L3. We outline our heuristic to reorder the ranking below:

1. Get top–5 L3 research fields, top–4 L2 research fields, and top–3 L1 research fields.
2. Assign initial score v for each research field based on its
   ranking.

   $$v(f_i) = (K - i)/K$$

   where K is the length of top-k, and i is the ranking of a research field f. For example, research fields in top–5 L3 have initial score of `[1, 0.8, 0.6, 0.4, 0.2]`, top–4 L2 have initial score of `[1, 0.75, 0.5, 0.25]`, and top–3 L1 have `[1, 0.666, 0.333]`
3. Update the score by multiplying each score with the score of
   matching research fields at upper level, and 0 otherwise.

   $$score(f_i^l) = \begin{cases} \prod_{l \in L} v(f^l) & \text{if field matched} \\ 0 & \text{otherwise} \end{cases}$$

   where L is the level of research field f and
   its upper levels. Here are examples of score update:
   - Soc–2–4 at rank–2 in L3, Soc–2 at rank–3 in L2, and Soc at
   rank–1 in L1. In this case, the score of Soc–2–4 is
   `0.8 * 0.5 * 1 = 0.4.`
   - Soc–2–4 at rank–1 in L3, Soc–2 at rank–2 in L2, but Soc is not found in top rank in L1. In this case, the score of Soc–2–4 is `0`.
4. Collect score from L2 and L3, and exclude L2 if we see more specific of L2 in top–5 L3.
5. Re-rank L2 and L3 research fields based on the score.
6. Return research fields having score $>= 0.4$.

To expand to more context from paper list in bibliography section, we also build other three Naive Bayes classifiers for L1, L2, and L3 using paper title feature only. We believe that a publication from a certain field also cites other publications from same or similar fields. For each publication in the bibliography, we apply the same procedure as mentioned above, then we average the score to get top research fields from bibliography. Finally, we combine top research fields from paper titles and abstract with results from bibliography.

Experiment Results

We discuss our experiment results for each task in this section. We use standard precision, recall, and F1 as evaluation metrics.

Dataset Extraction

First, we analyze our experiment for dataset detection subtask comparing Naive Bayes and SVM classifier. Using only paper titles in bibliography and explicit research method mentions, Naive Bayes and SVM classifiers are able to reach 0.88 & 0.92 F1 score respectively. Since SVM outperforms Naive Bayes, we use SVM for our dataset detection module. Table 10.4 shows detail dataset detection results on development set.

Table 10.4: Dataset Detection Results on Development Set

Classifier	Prec.	Rec.	F1
Naive Bayes	0.85	0.92	0.88
SVM	0.96	0.88	0.92

To see the impact of performing dataset detection, we test the performance of dataset extraction with and without dataset detection on development set. Table 10.5 summarizes the results. As shown in the table, performing dataset detection before extraction significantly improves the dataset extraction on development set.

Table 10.5: Dataset Extraction Results on Development Set

Method	Prec.	Rec.	F1
No Dataset Detection	0.18	0.33	0.24
With Dataset Detection	0.34	0.30	0.32

Table 10.6: Dataset Extraction Result on Test Set

Dataset	Prec.	Rec.	F1
Test Set (phase1)	0.17	0.10	0.13

Table 10.6 shows dataset extraction performance on test set (phase 1). The significant drop from development set result suggests that the test set might have different distribution compare to the training and development set. It might also contain dataset citations that are never been seen in training set. As mentioned in chapter 5, the test set contains new data source, non-open access journals from Sage publications which are not available in training and development set. It would be better to evaluate open access publications and non-open access publications separately so that we can have better understanding on the characteristics on each source in test set.

Research Methods Identification

We only consider Naive Bayes and Logistic Regression classifiers for research method identification because they naturally outputs probability score. We perform 5-fold cross validation to evaluate classification performance, and the result can be seen in table 10.7. Logistic regression classifier outperforms Naive Bayes with 0.86 F1 score in classifying 133 research methods.

Table 10.7: F1 Score for Research Method Classification

Classifier	F1
Naive Bayes	0.55
Logistic Regression	0.86

Research Fields Identification

We perform 5-fold cross validation to evaluate our classifiers to classify L1, L2, and L3 research fields. Table 10.8 shows the results using n-gram features from paper title and abstract, whereas table 10.9 shows the results using n-gram features from title only. Naive Bayes tends to perform slightly better on L3 research fields where we have large number of research field labels. We decide to use SVM for research field identification on publication level because SVM is generally better than Naive Bayes. On the other hand, we decide to use Naive Bayes for research field identification on bibliography level because Naive Bayes prefer to have more accurate L2 and L3 research fields.

Table 10.8: F1 Score for Research Field Classification on Publication Level using Paper Title and Abstract

Classifier	L1	L2	L3
Naive Bayes	0.78	0.37	0.13
SVM	0.82	0.38	0.12

Table 10.9: F1 Score for Research Field Classification on Bibliography Level using Paper Title Only

Classifier	L1	L2	L3
Naive Bayes	0.80	0.35	0.12
SVM	0.81	0.35	0.11

Lesson Learned

Extraction of research datasets, associated research methods and fields from social science publication is challenging, yet an important problem to organize social science publications. We have described our approach for the RCC challenge, and table 10.10 summarizes our approach. Beside publication content such as paper titles, abstract, full text, our approach also leverages on the information from bibliography. Furthermore, we also collect external information from SAGE Knowledge to get more information about research fields.

Table 10.10: Summary of Our Approach

Method	Features (n-gram)
Dataset extraction	
SVM for dataset detection	paper titles in bibliography and explicit research method mentions
Implicit entity linking	paper title and full text
Research method identification	
Logistic regression	paper title, abstract, and full text
Research field identification	
SVM (on paper)	paper title and abstract
Naive Bayes (on bibliography)	paper titles in bibliography

Apart from F1 score on 5-fold cross validation, we have no good way to evaluate research method and research field identification without ground truth label. Our methods are unable to automatically extract and recognize new datasets, research methods, and fields. An extension to automatically handle such cases using advance Natural Language Processing (NLP) approach is a promising future direction. All RCC finalists have shown that NLP approaches worked well on dataset extraction. Readers are encouraged to read their solutions on chapter 6, 7, 8, and 9.

Our model did not perform well in test set, and unable to advance to the second phase. Nevertheles, we recommend to use our approach as a baseline method as it is simple, efficient, and fast to train. From this competition, we have learned that lacks of labelled training data is a huge challenge, and it directs us to other external resources (i.e., SAGE Knowledge) as proxy for our label. We are also interested in exploring more advanced information extraction approaches on the RCC datasets. Another challenge is data sparsity. Although we see many papers listed in bibliography, lacks of access to these publication make us difficult to exploit citation network. Expanding labeled data from the list of papers in bibliography will be very beneficial to improve rich context of paper publications and datasets.

Acknowledgments

This research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its International Research Centres in Singapore Funding Initiative.

References

- [PM04] Fuchun Peng and Andrew McCallum (2004): Accurate information extraction from research papers using conditional random fields. In HLT-NAACL.
- [Het08] Erik Hetzner (2008): A simple method for citation metadata extraction using hidden markov models. In JCDL.
- [BREM12] Katarina Boland, Dominique Ritze, Kai Eckert, and Brigitte Mathiak (2012): Identifying references to datasets in publications. In TPDL.
- [APBM14] Sam Anzaroot, Alexandre Passos, David Belanger, and Andrew McCallum (2014): Learning soft linear constraints with application to citation field extraction. In ACL.
- [NCKL15] Viet Cuong Nguyen, Muthu Kumar Chandrasekaran, Min-Yen Kan, and Wee Sun Lee (2015): Scholarly document information extraction using extensible features for efficient higher order semi-crfs. In JCDL.
- [GML$^+$16] Behnam Ghavimi, Philipp Mayr, Christoph Lange, Sahar Vahdati, and Sören Auer (2016a): A semi-automatic approach for detecting dataset references in social science texts. Inf. Services and Use, 36:171–187.
- [GMVL16] Behnam Ghavimi, Philipp Mayr, Sahar Vahdati, and Christoph Lange (2016b): Identifying and improving dataset references in social sciences full texts. In ELPUB.
- [SBP$^+$16] Mayank Singh, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee (2016): Ocr++: A robust framework for information extraction from scholarly articles. In COLING.
- [ADR$^+$17] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum (2017): Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In SemEval@ACL.
- [AGJ$^+$17] Dong An, Liangcai Gao, Zhuoren Jiang, Runtao Liu, and Zhi Tang (2017): Citation metadata extraction via deep neural network-based segment sequence labeling. In CIKM.
- [AS17] Isabelle Augenstein and Anders Søgaard (2017): Multi-task learning of keyphrase boundary classification. In ACL.
- [LOH17] Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi (2017): Scientific information extraction with semi-supervised neural tagging. In EMNLP.
- [ACK18] Danny Rodrigues Alves, Giovanni Colavizza, and Frédéric Kaplan (2018): Deep reference mining from scholarly literature in the arts and humanities. In Front. Res. Metr. Anal.
- [AGB$^+$18] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni (2018): Construction of the literature graph in semantic scholar. In NAACL-HTL.
- [NJM18] Zara Nasar, S. W. Jaffry, and Muhammad Kamran Malik (2018): Information extraction from scientific articles: a survey. Scientometrics, 117:1931–1990.

Appendix: Technical Documentation

Source codes to run and replicate our experiments are available at https://github.com/LARC-CMU-SMU/coleridge-rich-context-larc.

Chapter 11 - Finding datasets in publications: The University of Syracuse approach

author:
- Tong Zeng\$$^{\{1}$,2\}$and Daniel Acuna^{\{1\}\$[}$1]
bibliography:
- ‘rcc–06.bib’
subtitle: |
Dataset mention extraction in scientific articles using a BiLSTM-CRF model
title: ‘Finding datasets in publications: The Syracuse University approach’

\$$^{\{1\}\$School}$ of Information Studies, Syracuse University, Syracuse, USA\
\$$^{\{2\}\$School}$ of Information Management, Nanjing University, Nanjing, China

Abstract

Datasets are critical for scientific research, playing a role in replication, reproducibility, and efficiency. Researchers have recently shown that datasets are becoming more important for science to function properly, even serving as artifacts of study themselves. However, citing datasets is not a common or standard practice in spite of recent efforts by data repositories and funding agencies. This greatly affects our ability to track their usage and importance. A potential solution to this problem is to automatically extract dataset mentions from scientific articles. In this work, we propose to achieve such extraction by using a neural network based on a BiLSTM-CRF architecture. Our method achieves $F_1 = 0.885$ in social science articles released as part of the Rich Context Dataset. We discuss future improvements to the model and applications beyond social sciences.

Introduction

Science is fundamentally an incremental discipline that depends on previous scientist's work. Datasets form an integral part of this process and therefore should be shared and cited as any other scientific output. This ideal is far from reality: the credit that datasets currently receive does not correspond to their actual usage[@datarank]. One of the issues is that there is no standard for citing datasets, and even if they are cited, they are not properly tracked by major scientific indices. Interestingly, while datasets are still used and mentioned in articles, we lack methods to extract such mentions and properly reconstruct dataset citations. The Rich Context Competition challenge aims at closing this gap by inviting scientists to produce automated dataset mention and linkage detection algorithms. In this article, we detail our proposal to solve the dataset mention step. Our approach attempts to provide a first approximation to better give credit and keep track of datasets and their usage.

The problem of dataset extraction has been explored before. @ghavimiIdentifyingImprovingDataset2016 and @ghavimiSemiautomaticApproachDetecting2017 use a relatively simple tf-idf representation with cosine similarity for matching dataset identification in social science articles. Their method consists of four major steps: preparing a curated dictionary of typical mention phrases, detecting dataset references, and ranking matching datasets based on cosine similarity of tf-idf representations. This approach achieved a relatively high performance, with $F_1 = 0.84$ for mention detection and $F_1 = 0.83$, for matching. @singhalDataExtractMining2013 proposed a method using normalized Google distance to screen whether a term is in a dataset. However, this method relies on external services and is not computational efficient. They achieve a good $F_1 = 0.85$ using Google search and $F_1 = 0.75$ using Bing. A somewhat similar project was proposed by @luDatasetSearchEngine2012. They built a dataset search engine by solving the two challenges: identification of the dataset and association to a URL. They build a dataset of 1000 documents with their URLs, containing 8922 words or abbreviations representing datasets. They also build a web-based interface. This shows the importance of dataset mention extraction and how several groups have tried to tackle the problem.

In this article, we describe a method for extracting dataset mentions based on a deep recurrent neural network. In particular, we used a Bidirectional Long short-term Memory (BiLSTM) sequence to sequence model paired with a Conditional Random Field (CRF) inference mechanism. The architecture is similar to chapter 6, but we only focus on the detection of dataset mentions. We tested our model on a novel dataset produced for the Rich Context Competition challenge. We achieve a relatively good performance of $F_1 = 0.885$. We discuss the limitations of our model.

The dataset

The Rich Context Dataset challenge was proposed by the New York University's Coleridge Initiative [@richtextcompetition]. The challenge comprised several phases, and participants moved through the phases depending on their performance. We only analyze data of the first phase. This phase contained a list of datasets and a labeled corpus of around 5K publications. Each publication was labeled indicating whether a dataset was mentioned within it and which part of the text mentioned it. The challenge used the accuracy for measuring the performance of the competitors and also the quality of the code, documentation, and efficiency.

We adopt the CoNLL 2003 format [@tjong2003introduction] to annotate whether a token is a part of dataset mention. Concretely, we use the tag DS denotes a dataset mention; The B- prefix indicates that the token is the beginning of a dataset mention, the I- prefix indicates the token is inside of dataset mention, and O denotes a token that is not a part of dataset mention. We put each token and its tag (separated by horizontal tab control character) in one line, and use the end of line (\n) control character as separator between sentences (see Table 2.1). The dataset were randomly split by 70%, 15%, 15% for training set, validation set and testing set, respectively.

Table 2.1. Example of a sentence annotated by IOB tagging format.

Token	Annotation
This	O
…	…
data	O
from	O
the	O
Monitoring	B-DS
the	I-DS
Future	I-DS
(	O
MTF	B-DS
)	O
\n	

The Proposed Method

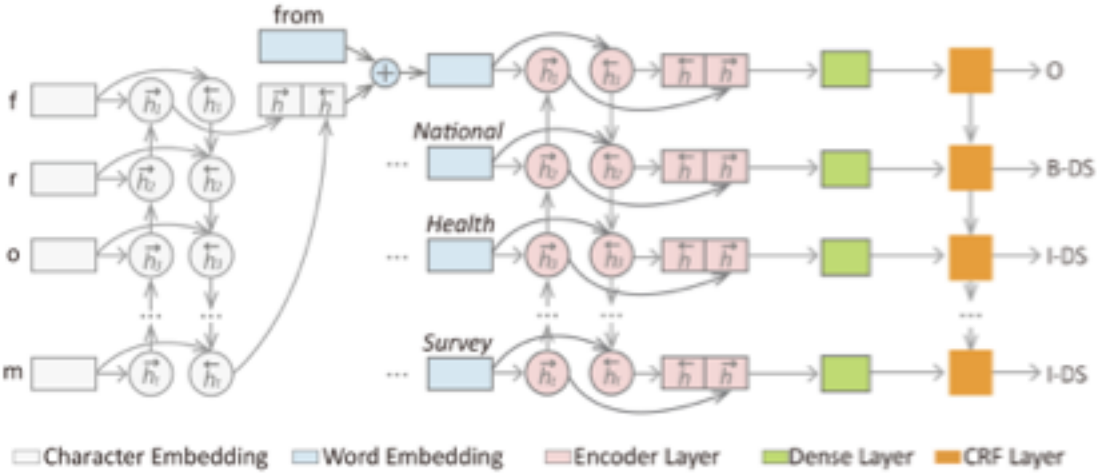
Overall view of the architecture

In this section, we propose a model for detecting mentions based on a BiLSTM-CRF architecture. At a high level, the model uses a sequence-to-sequence recurrent neural network that produces the probability of whether a token belongs to a dataset mention. The CRF layer takes those probabilities and estimates the most likely sequence based on constrains between label transitions (e.g., mention–to–no-mention–to-mention has low probability). While this is a standard architecture for modeling sequence labeling, the application to our particular dataset and problem is new.

We now describe in more detail the choices of word representation, hyper-parameters, and training parameters. A schematic view of the model is in Figure 3.1 and the components are as follows:

1. Character encoder layer: treat a token as a sequence of characters and encode the characters by using a bidirectional LSTM to get a vector representation.
2. Word embedding layer: mapping each token into fixed sized vector representation by using a pre-trained word vector.
3. BiLSTM layer: make use of Bidirectional LSTM network to capture the high level representation of the whole token sequence input.

4. Dense layer: project the output of the previous layer to a low dimensional vector representation of the the distribution of labels.
5. CRF layer: find the most likely sequence of labels.



{width="80%"}

Figure 3.1. Network Architecture of BiLSTM-CRF network

Character encoder

Similar to the bag of words assumption, a word could be composed of characters sampled from a bag of characters. Previous research [@santos2014learning; @jozefowicz2016exploring] has shown that the use of character-level embedding could benefit multiple NLP-related tasks. In order to use character-level information, we break down a word into a sequence of characters, then build a vocabulary of characters. We initialize the character embedding weights using the vocabulary size of a pre-defined embedding dimension, then update the weights during the training process to get the fixed-size character embedding. Next, we feed a sequence of the character embedding into an encoder (a bidirectional LSTM network) to produce a vector representation of a word. By using a character encoder, we can solve the out-of-vocabulary problem for pre-trained word embedding, as every word could be composed of characters.

Word Embedding

The word embedding layer is responsible for storing and retrieving the vector representation of words. Concretely, the word embedding layer contains a word embedding matrix $M^{tkn} \in \mathbb{R}^{|V|d}$, where the V is the vocabulary of the tokens and the d is the size of the embedding vector. The embedding matrix was initialized by a pre-trained GloVe vectors [@pennington2014glove], and updated by learning from the data. In order to retrieve from the embedding matrix, we first convert a given sentence into a sequence of tokens, then for each token we lookup the embedding matrix to get its vector representation. Finally, we get a sequence of vectors as input for the encoder layer.

LSTM

The Recurrent Neural Network (RNN) is a type of artificial neural network which takes the output of previous step as input of the current step recurrently. This recurrent nature allows it to learn from sequential data, for example, the text which consists of a sequence of works. RNN could capture contextual information in variable-length sequences in theory but it suffers from gradient exploding/vanishing problems [@pascanu2013difficulty]. The Long Short-Term Memory (LSTM) architecture was proposed by @hochreiter1997long to cope with these gradient problems. Similar to standard RNN, the LSTM network also has a repeating module called LSTM cell. The cell remembers information over arbitrary time steps because it allows information to flow along it without change. The cell state is regulated by a forget gate and an input gate which control the proportion of information to forget from a previous time step and to remember for a next time step. Also, there is a output gate controlling the information to flow out of the cell. The LSTM could be defined formally by the following equations:

$$i_t = \sigma(W_i x_t + W_i h_{t-1} + b_i)$$

$$f_t = \sigma(W_f x_t + W_f h_{t-1} + b_f)$$

$$g_t = tanh(W_g x_t + W_g h_{t-1} + b_g)$$

$$o_t = \sigma(W_o x_t + W_o h_{t-1} + b_o)$$

$$c_t = f_t \bigotimes c_{t-1} + i_t \bigotimes g_t$$

$$h_t = o_t \bigotimes tanh(c_t)$$

where x_t is the input at time t, W is the weights, b is the bias. The σ is the sigmoid function, $\bigotimes$ denotes the dot product, c_t is the LSTM cell state at time t and h_t is hidden state at time t. The i_t, f_t, o_t and g_t are named as input, forget, output and cell gates respectively.

LSTM can learn from the previous steps, which is the left context if we feed the sequence from left to right. However, the information in the right context is also important for some tasks. The bidirectional LSTM [@graves2013speech] satisfies this information need by using two LSTMs. Concretely, one LSTM layer was fed by a forward sequence and the other by a backward sequence. The final hidden states of each LSTM were concatenated to model the left and right contexts

$$h_t = [\overrightarrow{h_t} \backslash \mathrm{varoplus} \overleftarrow{h_t}]$$

Finally, the outcomes of the states are taken by a Conditional Random Field (CRF) layer that takes into account the transition nature of the beginning, intermediate, and ends of mentions. For a reference of CRF, refer to [@lafferty2001conditional]

Results

In this work, we wanted to propose a model for the Rich Context Competition challenge. We propose a relatively standard architecture based on a BiLSTM-CRF recurrent neural network. We now describe the evaluation metrics, hyper-parameter setting, and the results of this network on the dataset provided by the competition.

For all of our results, we use F_1 as the measure of performance. This measure is the harmonic average of the precision and recall and it is the standard measure used in sequence labeling tasks. This metric varies from 0 to 1, the higher the better. Our method achieved a relatively high F_1 of 0.885 for detecting mentions.

Table 4.1. Model search space and best assignments

Hyper-parameter	Search space	Best parameter
number of epochs	50	50
patience	10	10
batch size	64	64
pre-trained word vector size	choice[50, 100, 200,300]	100
encoder hidden size	300	300
number of encoder layers	2	2
dropout rate	choice[0.0,0.5]	0.5
learning rate optimizer	adam	adam
l2 regularizer	0.01	0.01
learning rate	0.001	0.001

We train models using the training data and monitor the performance using the validation data (we stop training if the performance does not improve for the last 10 epochs). We are using the Adam optimizer with learning rate of 0.001 and batch size equal to 64. The hidden size of LSTM for character and word embedding is 80 and 300, respectively. For the regularization methods, and to avoid over-fitting, we use L2 regularization set to 0.01 and we also use dropout rate equal to 0.5. We trained 8 models with a combination of different GloVe vector size (50, 100, 300 and 300) and dropout rate (0.0, 0.5). The hyper-parameter settings are present in Table 4.1.

Table 4.2. Performance of proposed network

Models	GloVe size	Dropout rate	Precision	Recall	F1
m1	50	0.0	0.884	0.873	0.878
m2	50	0.5	0.877	0.888	0.882
m3	100	0.0	0.882	0.871	0.876
m4	100	0.5	0.885	0.885	0.885
m5	200	0.0	0.882	0.884	0.883
m6	200	0.5	0.885	0.880	0.882
m7	300	0.0	0.868	0.886	0.877
m8	300	0.5	0.876	0.878	0.877

The test performances are reported in Table 4.2. The best model is trained by word vector size 100 and dropout rate 0.5 with F_1 score 0.885 (Table 4.2), and it takes 15 hours 58 minutes for the training on an NVIDIA GTX 1080 Ti GPU in a computer with an Intel Xeon E5–1650v4 3.6 GHz CPU with 128 GB of RAM.

We also found some limitations to the dataset. Firstly, we found that mentions are nested (e.g. HRS, RAND HRS, RAND HRS DATA are linked to the same dataset). The second issue most of the mentions have ambiguous relationships to datasets. In particular, only 17,267 (16.99%) mentions are linked to one dataset, 15,292 (15.04%) mentions are listed to two datasets, and 12,624 (12.42%) are linked to three datasets. If these difficulties are not overcome, then the predictions from the linkage process will be noisy and therefore impossible to tell apart.

Conclusion

In this work, we report a high accuracy model for the problem of detecting dataset mentions. Because our method is based on a standard BiLSTM-CRF architecture, we expect that updating our model with recent developments in neural networks would only benefit our results. We also provide some evidence of how difficult we believe the linkage step of the challenge could be if the dataset noise are not lowered.

One of the shortcomings of our approach is that the architecture is lacking some modern features of RNN networks. In particular, recent work has shown that attention mechanisms are important especially when the task requires spatially distant information, such as this one. These benefits could also translate to better linkage. We are exploring new architectures using self-attention and multiple-head attention. We hope to explore these approaches in the near future.

There are number of improvements that we can make in the future. A first improvement is to use non-recurrent neural architectures such as the Transformer which has shown to be faster and a more effective learner compared to recurrent neural networks. Another improvement would be to bootstrap information from other dataset sources such as open access full-text articles from PubMed Open Access Subset. This dataset contains dataset *citations* [@datarank]—in contrast to the most common types of citations to publications. The location of this citations within the full-text could be exploited to perform entity recognition. While this would be a somewhat different problem than the one solved in this article, it would still be useful for the goal of tracking dataset usage. In sum, by improving the learning techniques and the dataset size and quality, we could significantly increase the success of finding datasets in publications.

Our proposal, however, is surprisingly effective. Because we have barely modified a general RNN architecture, we expect that our results will generalize relatively well either to the second phase of the challenge or even to other disciplines. We would emphasize, however, that the quality of the dataset has a great deal of room for improvement. Given how important this task is for the whole of science, we should try to strive to improve the quality of these datasets so that techniques like this one can be more broadly applied. The importance of dataset mention and linkage therefore could be fully appreciated by the community.

Acknowledgements {#acknowledgements .unnumbered .unnumbered}

Tong Zeng was funded by the China Scholarship Council #201706190067. Daniel E. Acuna was funded by the National Science Foundation awards #1646763 and #1800956.

\bibliographystyle{apalike}

Chapter 12 - The future of context

The Future of AI in Rich Context

Paco Nathan

The setting for Rich Context originated in needs to analyze confidential micro-data for evidence-based policymaking.
The nature of that work involves collaboration using *linked data*, with strict requirements for data privacy and security, plus provisions for data stewardship and dataset curation.
Due to the highly regulated environments, that data cannot be examined outside of its specific use cases.
In a world where public search engines crawl and index millions of terabytes, making search results available within milliseconds to anyone with a browser and an Internet connection, the setting for Rich Context may appear utterly alien.
Seemingly, a reasonable compromise would be to run queries of sensistive data within their secure environments, and otherwise leave the process unexamined.

However, there’s a broader scope to consider, far beyond the process of managing research projects or curating particular datasets.
The great challenges of our time are human in nature – climate change, terrorism, overuse of natural resources, the nature of work, and so on – and these require robust social science to understand their causes and consequences.
Effective use of data for social science research depends on understanding how datasets have been produced and how they’ve been used in previous works.
That understanding of data provenance is complicated by the fact that research often must link datasets from different data producers, different agencies, different organizations.

Other factors confound this situation.
On the one hand, the availability of inexpensive computing resources, ubiquitous connected mobile devices, social networks with global reach, etc., implies that researchers can acquire large, rich datasets.
Researchers can also fit statistical models that might have seemed intractably complex merely a decade ago.
On the other hand, accumulating important information about datasets, their provenance, and their usage has historically been a manual process.
Sharing this kind of information across organizations is difficult in general, and when datasets include confidential data about human subjects it becomes impossible to provide open access to the original data.
These issues combine to contribute to a lack of reproducibility and replicability in the study of human behavior, and threaten the legitimacy and utility of social science research.

The problems enumerated above make it difficult for people to understand about data usage, although at the same time they present opportunities for leveraging automation.
Consider how one of the major challenges in social science research is search and discovery: the vast majority of data and research results cannot be easily discovered by other researchers.
From one perspective, researchers are the users of micro data and its related *metadata* – in other words, information about the structure of datasets, their provenance, etc. – and those researchers produce outcomes, often in the form of publications.
Publications accumulate expertise and nuances about datasets, including the data preparation required, research topics and methodology, what kinds of analyses were attempted and ultimately used, which information within the datasets was most valuable for the results obtained, and so on.
These details produced through publications represent *metadata about datasets*.
While the metadata within publications may be relatively unstructured – i.e., not explicitly articulated, nor shared outside of the current project – advances in machine learning provide means to extract metadata from unstructured sources.

The exchange of metadata plays another important role.
From the perspective of a data publisher (i.e., an agency) the many concerns about security and data privacy indicate use of *tiered access* for sensitive data.
Datasets which do not contain sensitive data may be made freely available to the public as *open data*.
Other datasets may require DUAs before researchers can access them.
So the data sharing may need to be organized in tiers.
Nonetheless, metadata for the private tiers in many cases may still be shared even though the data cannot be linked directly without explicit authorizations and stewardship.
So metadata provides a role of exchanging information about sensitive data, in ways that can be accumulated across a broader scope than individual research projects.

The opportunity at hand is to leverage machine learning advances to create feedback loops among the entities involved: researchers, datasets, data publishers, publications, and so on.
A new generation of tooling for search and discovery could leverage that to augment researchers: informing them about what datasets are being used, in which research fields, the tools involved, as well as the methods used and findings from the research.

The Case for Rich Context

Consider the two most fundamental workflows within Rich Context, where analysts and other researchers interact among data providers, data stewards, training programs, security audits, etc.:

- **Collaboration and Workspace:** where researchers collaborate within a secured environment, having obtained authorizations via NDAs (non-disclosure agreements), DUAs (data use authorizations), etc.
- **Data Stewardship:** where data stewards can review and determine whether to approve requests for using the datasets that they curate, and then monitor and report on subsequent usage.

These components represent *explicitly* linked feedback loops among the researchers, projects, datasets, and data stewards.
Researchers also use other *implicitly* linked feedback loops externally to draw from published social science research.
Overall, the general category of linked data describes these interactions.

A large body of AI applications leverages linked data.
Related R&D efforts have focused mostly on public search engines, e-commerce platforms, and research in life sciences – while in social science research the use of this technology is relatively uncharted territory.
Also, given the security and compliance requirements involved with sensitive data, the process of leveraging linked data in social science research takes on nuanced considerations and compels novel solutions.

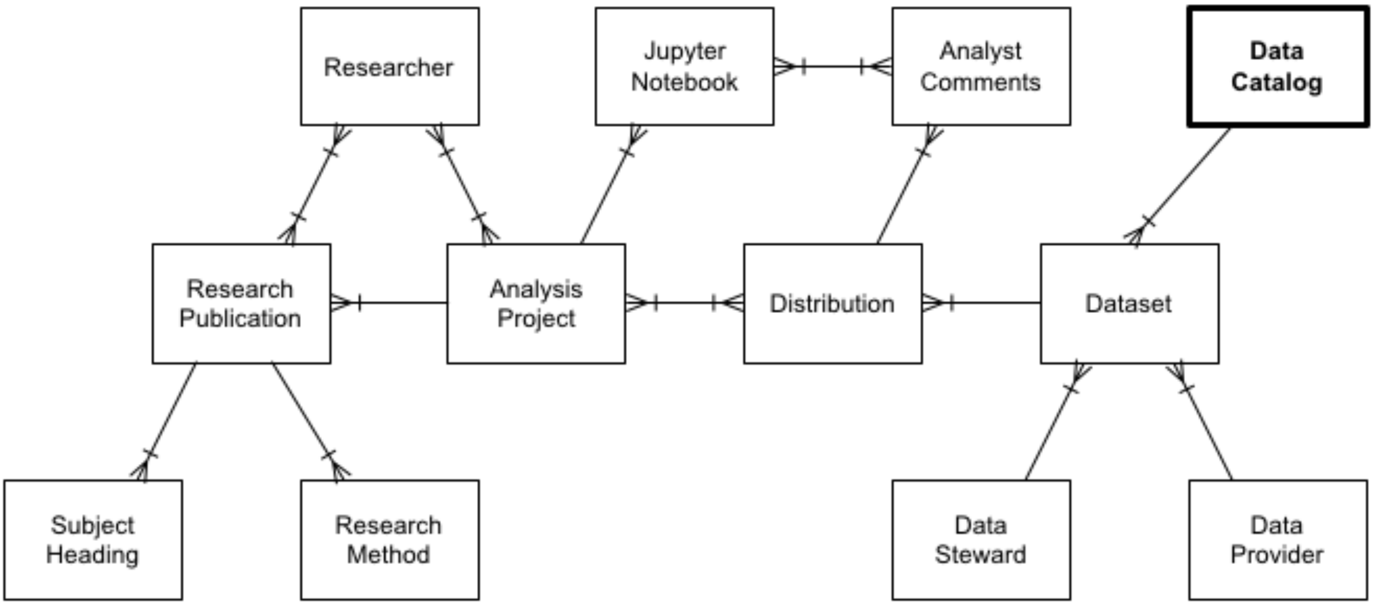
This area of focus represents the core of Rich Context: the interconnection of point solutions that facilitate research, as explicit feedback loops, along with means to leverage the implicit feedback loops that draw from published research.
Making use of AI applications to augment social science research is the goal of Rich Context work, and that interconnection of feedback loops, through a graph, creates a kind of *virtuous cycle* for metadata – analogous to the famous virtuous cycle of data required for AI applications in industry, as described Andrew Ng.

In general, guidance for Rich Context can be drawn from the FAIR[1] data principles for data management and data stewardship in science.
The FAIR acronym stands for *Findable*, *Accessible*, *Interoperable*, and *Reusable* data, addressing the issue of reproducibility in scientific research.
One observation from the original FAIR paper describes core tenets of Rich Context:

> *Humans, however, are not the only critical stakeholders in the milieu of scientific data. Similar problems are encountered by the applications and computational agents that we task to undertake data retrieval and analysis on our behalf. These 'computational stakeholders' are increasingly relevant, and demand as much, or more, attention as their importance grows. One of the grand challenges of data-intensive science, therefore, is to improve knowledge discovery through assisting both humans, and their computational agents, in the discovery of, access to, and integration and analysis of, task-appropriate scientific data and other scholarly digital objects.*

In other words, throughout the use cases for scientific data there are substantial opportunities for human-in-the-loop AI approaches, where the people involved increasingly have their work augmented by automated means, while the automation involved increasingly gets improved by incorporating human expertise.
One can use the metaphor of a *graph* to represent the linkages: those that span across distinct research projects, those that require cross-agency collaboration with sensitive data, and those that integrate workflows beyond the scope of specific tools.

Specifically, this work entails the development of a *knowledge graph* to represent metadata about datasets, researchers, projects, agencies, etc., – including the computational agents involved – as distinct entities connected through relations that model their linkage.



example graph relations

Much of the intelligence in this kind of system is based on leveraging inference across the graph, insights which could not be inferred within the scope of a single research project or through the use of one particular tool.
Over time, the process accumulates a richer context of relations into that graph while clarifying and leveraging the feedback loops among the entities within the graph.
Rich Context establishes foundations for that work in social science research.

The Rich Context Competition held during September 2018 through February 2019 invited AI research teams to compete in one aspect of Rich Context requirements.
Several teams submitted solutions to automate the discovery of research datasets along with associated research methods and fields, as identified in social science research publications.
Methods for machine learning and text analytics used by the four finalist teams provided complementary approaches, all focused on the problem of *entity linking*, with a corpus of social science research papers used as their training data.

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.
There are many publication sources to analyze, and the project will pursue that work as an ongoing process to extract the implied metadata.
Meanwhile the increasing adoption and usage of the ADRF framework continues to accumulate metadata directly.

Use Cases for Rich Context

Looking at potential use cases for Rich Context more formally, we can identify needs for leveraging a knowledge graph about research datasets and related entities.
For each of these needs, we can associate solutions based on open source software which have well-known use cases in industry.

As an example, consider a dataset *A001* published by a data provider *XYZ Agency* where *Jane Smith* works as the data steward responsible for curating that dataset.
Over time, multiple research projects describe the use of *A001* in their published results.
Some researchers note, on the one hand, that particular columns in data tables within *A001* have some troubling data quality issues – inconsistent names and acronyms, identifiers that require transformations before they can be used to join with other datasets, and so on.
On the other hand, the body of research related to *A001* illustrates how it gets joined frequently with another dataset *B023* to support analysis using a particular research method.
The two datasets provide more benefits when used together.

While access to the *A001* dataset gets managed through the *XYZ Agency* and its use of the ADRF framework, other datasets such as *B023* get used outside of that context.
A knowledge graph is used to accumulate information about the datasets, research projects, the resulting published papers, etc., and applications for augmenting research derive quite directly from that graph.
For example, feedback from researchers about how *A001* gets combined with other datasets outside of the *XYZ Agency* domain help guide *Jane Smith* to resolve some of the data quality issues.
New columns get added with cleaner data for identifiers, which allows more effective linking.
Other feedback based on machine learning models that have classified published papers then helps recommend research methods and candidate datasets to new analysts – and also to agencies that have adjacent needs, but did not previously have visibility into the datasets published by *XYZ Agency*.

These are the kinds of applications that become enabled through Rich Context.
Search and discovery is clearly a need, although other use cases can help improve the discovery process and enhance social science research.
The following sections discuss specific use cases and their high-level requirements for the associated technologies.

Search and Discovery

As described above, the vast majority of social science data and research results cannot be easily discovered by other researchers.
While public search engines based on keyword search have been popularized by e-commerce platforms such as Google and Bing, the more general problem of search and discovery can be understood best as a graph problem, and the needs in social science research are more formally understood as recommendations across a graph.

For example, starting with a given dataset, who else has worked with that data?
Which topics did they research?
Which methods did they use?
What were their results?
In other words, starting from one entity in a knowledge graph, what other neighboring entities are linked?

These kinds of capabilities may be implemented simply by users traversing directly through the links of the graph.
However, at scale, that volume of information can become tedious and overwhelming.
It's generally more effective for user experience (UX) to have machine learning models summarize, then predict a set of the most likely paths through the graph from a particular starting point.

One good approach for this is the general case of *link prediction*[13]: given a researcher starting with a particular dataset and goals for topics or methods, represent that as a local, smaller graph.
Then use link prediction to fill-in missing entities and relations, extending the local graph for that researcher.
In other words, what other datasets should be joined, how can particular fields be used, what research topics or methods are related, which published papers might become foundations for this work?
The most likely links inferred become top recommendations.
Also, this kind of recommendation is not limited to the start of projects, it can be leveraged at almost any stage of research.

Entity Linking

The Rich Context Competition demonstrated how entities and relations used to construct a knowledge graph can be mined from a corpus of scientific papers.
Machine learning methods for *entity linking*[14] used in the competition need to be generalized and extended, then used to analyze the ongoing stream of published social science research.
This work provides potential benefits for the publishers, for example helping them analyze and annotate newly published papers, developing dashboards about data impact metrics for journals or authors, and so on.

An additional benefit of entity linking is to help correct abbreviations, localized acronyms, or mistakes in linked data references.
This is an iterative process which will need integration and feedback with data stewardship workflows.

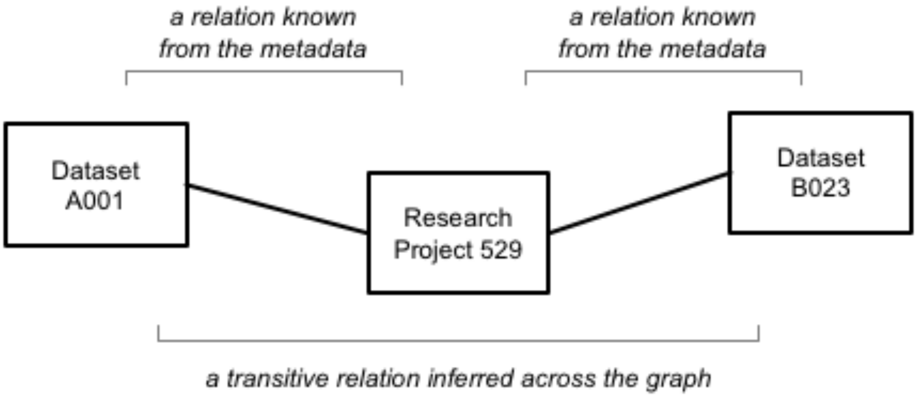
Classifiers

As any researcher or librarian knows well, curating a large set of research papers by hand is labor-intensive and prone to errors.
Machine learning models based on *supervised learning* or *semi-supervised learning* (human-in-the-loop) can produce classifiers that annotate research papers automatically.

At some point, the ADRF framework may run classifiers on the workflows (e.g., Jupyter notebooks) for projects in progress.
By extension, classifiers may infer across the knowledge graph to add annotations for datasets as well.
This work can be considered a subset of link prediction, also related to entity linking.

Transitive Inference

The metadata collected through the use of the ADRF framework or extracted from research publications includes relations that link entities in the graph.
Once a graph is constructed, additional relations may be inferred.
This is a case of *transitive inference*, which can help add useful annotations to the graph, as shown in the following diagram:


a relation known from the metadata
a relation known from the metadata
Dataset A001
Research Project 529
Dataset B023
a transitive relation inferred across the graph

transitive relations

In an example from Norse mythology, Torunn is the daughter of Thor, and Thor is the son of Gaea, therefore Torun is the *granddaughter* of Gaea.
The same process can apply, for example, to relations that describe links between datasets and

researchers.

Note that embeddings have proven to be a powerful approach for inference about patterns, based on deep learning.
On the current forefront of AI research, methods that leverage *reinforcement learning*[15] are positioned to outperform embeddings soon, since they explore/exploit the graph structure instead of relying on a history of observed patterns.
This is especially useful for *knowledge graph completion*, where there are cases of incomplete metadata in the knowledge graph, which is essential for Rich Context work.

Iterative Improvement of the Knowledge Graph

Most of the finalist teams in the Rich Context Competition made use of other existing graphs to bootstrap their machine learning development work, such as the Microsoft Academic Graph, Semantic Scholar, and others purpose-built for the competition.
Those teams cited how some graph would need to be extended in the future, to improve recognition accuracy.

Rich Context now subsumes that effort, making the iterative improvement of the knowledge graph an ongoing priority.
In lieu of those other graphs used for bootstrap purposes during the competition, the Rich Context knowledge graph provides the foundation for machine learning.

This process of accreting more entities into the graph and refining their relations leads to better training data and improved machine learning models.
Over time, as our models improve, the previously analyzed research papers can be re-evaluated to extract richer results.
That work in turn enhances social science research within the ADRF framework, along with data curation.
That overall dynamic represents the virtuous cycle of metadata, which continually improves Rich Context.

Axioms for Dataset Curation

Another immediate use of Rich Context is to assist the data stewards to understand the broader scope of usage for the datasets that they curate.
For example, ontology *axioms* used on the metadata in the graph can help analyze:

- consistency checks for the incoming metadata
- which data stewardship rules apply in a given case

In a way, that helps codify what would otherwise be “institutional lore” – instead that’s now captured for others to study, use for training new staff, etc.

Note that the ADRF framework must provide means for customizing and configuring these kinds of axioms, so that data stewardship rules rules are not tightly coupled with the security audits and release cycle.
Those rules can change rapidly, depending on new legislation or other policy updates, or even due to different agency environments.

Leveraging Open Standards and Open Source

Overall, the Rich Context portion of the ARDF framework represents a *data catalog* along with associated *data governance* practices.
As a first step in knowledge graph work, we can make use of existing open standards for metadata about data catalogs and datasets.
For example, the W3C Data Activity coordinates a wide range of metadata standards, including:

- DCAT – metadata about data catalogs
- VoID – metadata about datasets
- DCMI – Dublin Core metadata terms
- SKOS – “simple knowledge organization system”

These represent controlled vocabularies described in OWL and based atop RDF.
These standards can be combined and extended to suit the needs of specific use cases, such as within the ADRF framework.
In particular, the Rich Context knowledge graph is a superset of a *DCAT-compliant data catalog*.
Taken together, localized extensions of these open standards represent an ontology – essentially as a specification for defining metadata that can be added into the knowledge graph and how that graph should be structured.
Development of that ontology along with example metadata plus Python code to validate the graph is managed in the public repository adrf-onto on GitHub.

The workflows within the ADRF framework represent use cases of data governance, and there is substantial overlap between Rich Context and emerging trends for data governance in industry.
There are open source projects which leverage knowledge graphs to collect metadata about datasets and their usage, where machine learning helps address the complexities[16] of data governance in industry data science work.
For instance:

- Amundsen from Lyft
- Marquez from WeWork
- WhereHows from LinkedIn
- Databook from Uber (pending release as open source)

Of course the Rich Context work addresses special considerations for sensitive data and compliance requirements.
Even so, much can be learned from these related open source projects in industry, which are pursuing similar kinds of use cases.
TopQuadrant and AstraZeneca are examples of commercial vendors which construct knowledge graphs about datasets, also for data governance purposes – respectively in the Finance and Pharma business verticals.
These commercial solutions similarly make use of DCAT, VoID, DMCI, SKOS, and also the FAIR data principles.

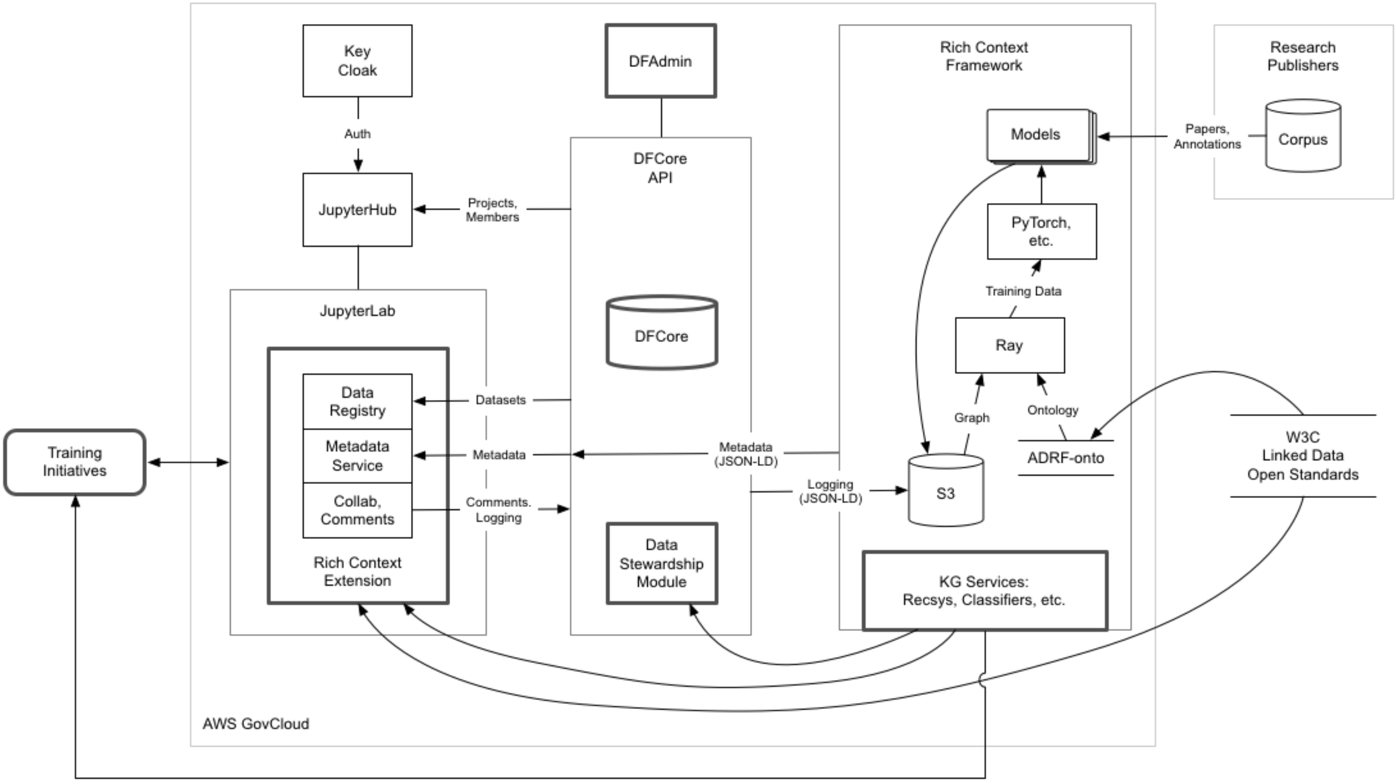
In general, the subject of metadata exchange for data governance use cases is addressed by the ODPi open standard Egeria and related work by Mandy Chessell[17], et al., including the Apache Atlas open source project.
Much of that work focuses on standards used to validate the exchange of metadata reliably across different frameworks.
This implies potential opportunities for Rich Context to interoperate with other data governance solutions or related metadata services.

To help establish open standards and open source implementations related to Rich Context, the ADRF team has collaborated with Project Jupyter.
A new Rich Context feature set is being added to JupyterLab, which is one of the key open source projects used in the architecture of the ADRF framework, and these new features will be integrated into its future releases.
The new Rich Context features support projects as top-level entities, real-time collaboration and commenting, data registry, metadata handling, annotations, and usage tracking – as described in the Project Jupyter “press release” requests for comments: data explorer, metadata explorer, and commenting.
For example, a team of social science researchers working on a project could use the commenting feature in Jupyter to make an annotation about data quality issues encountered in a particular dataset.
That comment, as metadata about the dataset, would get imported into the knowledge graph, and could later be used for recommendations to a data steward or other researchers.

Note that most of the machine learning approaches referenced above are specific cases of *deep learning*, based on layered structures of artificial neural networks. In particular, *graph embedding*[18] is an approach that vectorizes portions of graphs to use as training data for deep learning models. Graph embedding can be used to perform entity linking, link prediction, etc.
In many of these cases, the resulting machine learning models become proxies for the graph data, such that the entire knowledge graph data is not required in production use cases.
That practice contrasts earlier and generally less effective approaches which relied on graph queries applied to the full data.
Note that the winning team in the Rich Context Competition was from Allen AI 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 (from Facebook) and the more recent Ray (from UC Berkeley RISElab).

System Architecture Overview

The following diagram illustrates a proposed system architecture for Rich Context as an additional module in the ADRF framework:



Rich Context module

Building on the DFCore features plus the Data Stewardship module, Rich Context provides both a destination for metadata (logging events from components, or extracted metadata from analysis of publications) and a source for metadata ontology used in the ADRF framework.
Machine learning models get trained and updated based on the knowledge graph, then used for services (recommender system, classifiers, etc.) provided back into the ADRF framework, and additionally to support training initiatives – or for general purpose search and discovery by researchers.

The additional system components for implementing Rich Context are based primarily on open source software (e.g., PyTorch) and extensions of open standards (e.g., W3C), all within the security context of AWS GovCloud implementation of the ADRF framework.

Trends, from origins to near-term future projections

Meanwhile, the development of Rich Context has followed a familiar progression, echoing how the history of IT and data analytics practices matured over decades – albeit at a much faster pace. That progression indicates likely directions for how AI applications will come into use for Rich Context.
Initial steps for Rich Context allowed researchers to analyze and report about sensistive data, while maintaining security and privacy compliance.
That's roughly analogous to data analytics during the heyday of *enterprise data warehouses* and *business intelligence* during the 1990s.
Subsequent work improved online workflows for data stewards, adding reports about usage along with some metadata derived as "exhaust" from logs.
That's roughly analogous to "data driven" organizations that emerged during the late 2000s after initial adoption of *data science* practices.
Next steps, such as the Rich Context Competition, began to use *machine learning models* to extract metadata that was embedded in unstructured data (i.e., research publication) to augment research efforts.
That's roughly analogous to the trends of machine learning adoption in industry during the mid 2010s.
In the immediate future, Rich Context applications begin to leverage inference based on *knowledge graph* representations about researchers, datasets, publications, data stewards, and so on. Contemporary work in *deep learning* promises AI-based applications that can leverage embeddings in the graph, to give social science researchers better recommendations for their work.
While that depends on historical patterns, current research on *reinforcement learning* to explore/exploit the structure of graphs can move beyond history and patterns, effectively considering "what if" scenarios that suggest unexplored research opportunities.
That echoes the contemporary AI landscape, leading into the 2020s.

Summary

Rich Context recognizes that social science research depends on *linked data* usage of micro data and its metadata.
Effective management of that metadata is based on a graph that exists outside the context of component point solutions and specific workflows.
While there is substantial use of linked data for ecommerce platforms and research in life sciences, social science research presents nuances and new challenges that haven't been addressed previously.

The Rich Context portions of the ADRF framework interconnect workflows that facilitate research – as explicit feedback loops in the graph – along with means to extract metadata from published research – as implicit feedback loops in the graph.

That process creates a kind of virtuous cycle for metadata, making use of AI applications to augment social science research, with continual improvement of the entities and relations represented within the graph.

A prerequisite was to create a corpus of research publications, used for training data during the Rich Context Competition, which demonstrated how to extract metadata from research publications.

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. It will eventually incorporate the new Rich Context features going into Project Jupyter.
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 through the ADRF framework.

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.

Notes

[1]:

Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

[13]:

For a sample of recent research papers regarding link prediction through graph embedding, see these Arxiv results.

[14]:

One of the better resources online for entity linking is NLP-progress which specifically tracks the state-of-the-art (SOTA) papers, along with their scores on recognized benchmarks.

[15]:

between approaches based on RL vs. embedding, see “Multi-Hop Knowledge Graph Reasoning with Reward Shaping”; Xi Victoria Lin, Richard Socher, Caiming Xiong; *EMNLP 2018* arXiv:1808.10568 [cs.AI]

[16]:

A good survey paper about these issues is given Ground: A Data Context Service, Hellerstein et al., *CIDR 2017*, based on research by UC Berkeley RISElab.

[17]:

See “The Case for Open Metadata”, Mandy Chessell, *Frontiers in Data Science*, 2016–09–15.

[18]:

For an overview of graph embedding, see “Graph Embedding for Deep Learning”, Flawson Tong (2019–05–06).

1. Corresponding author: deacuna\@syr.edu ↩
2. The authors would like to thank Jannick Blaschke, Rafael Beier, and the editors for helpful comments. We would like to thank Jannick Blaschke for providing graphics. The views expressed here represent the authors‘ personal opinions and do not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem. ↩

Abstract

3. See Desai et al. (2016) for a detailed description of the Five Safes framework. ↩
4. Data producers in different departments across Bundesbank compile data, e.g. microdata, indicators, or time series. ↩
5. In our model, we have called this knowledge user specific knowledge. Here the knowledge is in that sense specific, that it can be used to fulfil the task of Bundesbank in a better way ↩
6. Chapter 4 discusses in more detail how metadata may support the discovery of microdata. ↩
7. See Chapter 5 for a more detailed description of the evaluation process used in the competition. ↩
8. footnote6 ↩
9. footnote6 ↩
10. footnote7 ↩
11. footnote7 ↩
12. For the future, ongoing effort is needed to support all four “corners of the circle” / “pillars of the level model”. The current competition strengthens the arrow from publication to knowledge and structures gained knowledge to improve data services. To support the data services pillar – for example -, digital RDC environments with facilitated access processes like a “data stewardship module” will in our view improve data access. ↩
13. https://www.ncbi.nlm.nih.gov/pmc/ ↩
14. https://www.aclweb.org/anthology/ ↩
15. https://github.com/allenai/science-parse ↩
16. [poppler]https://manpages.debian.org/testing/poppler-utils ↩
17. https://github.com/explosion/spaCy ↩
18. https://wiki.dbpedia.org/services-resources/ontology ↩
19. https://en.wikipedia.org/wiki/Category:Statistical_methods ↩
20. https://spacy.io/api/tokenizer ↩
21. https://rasa.com/docs/nlu ↩
22. https://nlp.stanford.edu/projects/glove ↩
23. https://fasttext.cc/docs/en/crawl-vectors.html ↩
24. https://www.w3.org/TR/shacl/ ↩
25. https://joinup.ec.europa.eu/release/dcat-ap/11 ↩
26. This differs from library or archival collections, which are usually thematically related, and for which the selection of items for inclusion is defined by an express collection policy. ↩
27. The challenges of metadata for data streams are related to the cataloguing of different editions of a work and of serials in a text-based library. ↩

28. https://wikidata.org/ ↩
29. https://duraspace.org/vivo/about/ ↩
30. https://ddialliance.org/Specification/RDF/XKOS ↩
31. http://www.obofoundry.org/ ↩
32. https://www.icpsr.umich.edu/ ↩
33. http://ddialliance.org ↩
34. DDI is also used for datasets from other organizations such as the National Opinion Research Center (NORC). ↩
35. https://www.cessda.eu/ ↩
36. https://www.europeansocialsurvey.org/data/ ↩
37. Australia National Data Service: https://www.ands.org.au/ ↩
38. There are additional collections at http://data.census.gov, http://gss.norc.org. [ ]{.underline} http://electionstudies.org, http://psidonline.isr.umich.edu, and http://www.nlsinfo.org. ↩
39. https://www.earthcube.org/ ↩
40. https://pds.nasa.gov/ ↩
41. https://www.uniprot.org/ ↩
42. http://www.rcsb.org/ ↩
43. re3data.org ↩
44. https://datadryad.org/ ↩
45. https://datacite.org/ ↩
46. https://www.crossref.org/ ↩
47. Health Level Seven International, https://www.hl7.org/ ↩
48. Kyoto Encyclopedia of Genes and Genomes, https://www.genome.jp/kegg/ ↩
49. See also, e.g., http://www.dcc.ac.uk/digital-curation/planning-preservation ↩
50. Controlled Lots of Copies Keep Stuff Safe, https://clockss.org/ ↩
51. http://irods.org ↩
52. The policies are based on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) standard. https://interparestrust.org/ ↩
53. http://www.dcc.ac.uk/sites/default/files/DRAMBORA_Interactive_Manual%5B1%5D.pdf; see also, http://www.dcc.ac.uk/resources/repository-audit-and-assessment/drambora. ↩
54. http://www.loc.gov/standards/premis/ontology/ ↩
55. https://www.w3.org/TR/prov-o/ ↩
56. The term micro-data is used in two distinct ways. In the context of HTML, it is associated with embedding Schema.org codes into web pages similar to micro-formats. In the context of survey data, it refers to individual-level data. ↩
57. https://jupyter.org/ ↩
58. http://wiss-ki.eu ↩
59. https://www.colectica.com/ ↩
60. https://gssdataexplorer.norc.org/ ↩
61. https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine–62d27254fbb9 ↩

62. https://mmisw.org/ ↩
63. https://www.w3.org/TR/vocab-data-cube/ ↩
64. https://www.earthcube.org/info/about ↩
65. http://sdmx.org/ ↩
66. https://taverna.incubator.apache.org/ ↩
67. https://www.myexperiment.org/about ↩
68. https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model. GSIM is coordinated with the Common Statistical Production Architecture (CSPA). https://unstats.un.org/unsd/nationalaccount/workshops/2015/gabon/BD/CSPA-ENG.pdf ↩
69. https://www.niem.gov/about-niem ↩
70. https://www.openarchives.org/pmh/ ↩