

EDITED BY

JULIA I. LANE • IAN MULVANY • PACO NATHAN



RICH SEARCH AND DISCOVERY FOR RESEARCH DATASETS

BUILDING THE NEXT GENERATION OF
SCHOLARLY INFRASTRUCTURE



RICH SEARCH AND DISCOVERY FOR RESEARCH DATASETS

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

EDITED BY

JULIA I. LANE • IAN MULVANY • PACO NATHAN

RICH SEARCH AND DISCOVERY FOR RESEARCH DATASETS

BUILDING THE NEXT GENERATION OF
SCHOLARLY INFRASTRUCTURE



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne



Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Jai Seaman
Assistant editor: Charlotte Bush
Assistant editor, digital: Sunita Patel
Production editor: Ian Antcliff
Copyeditor: Richard Leigh
Proofreader: Emily Ayers
Indexer: Martin Hargreaves
Marketing manager: Susheel Gokarakonda
Cover design: Shaun Mercier
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in the UK

Editorial arrangement © Julia I. Lane, Ian Mulvany and Paco Nathan 2020
Chapter 1 © Ian Mulvany, Paco Nathan, Sophie Rand and Julia Lane 2020
Chapter 2 © Stefan Bender, Hendrik Christian Doll and Christian Hirsch 2020
Chapter 3 © Christian Herzog, Daniel W. Hook, Mark Hahnel, Stacy Konkiel and Duane E. Williams 2020
Chapter 4 © Robert B. Allen 2020
Chapter 5 © Andrew Gordon, Ekaterina Levitskaya, Jonathan Morgan, Paco Nathan, Sophie Rand, Sebastian Kohlmeier and Alex D. Wade 2020
Chapter 6 © Daniel King, Waleed Ammar, Iz Beltagy, Christine Betts, Suchin Gururangan and Madeleine van Zuylen 2020
Chapter 7 © Haritz Puerto San Roman, Giwon Hong, Minh-Son Cao and Sung-Hyon Myaeng
Chapter 8 © Wolfgang Otto, Andrea Zielinski, Behnam Ghavimi, Dimitar Dimitrov, Narges Tavakolpoursaleh, Karam Abdulahhad, Katarina Boland and Stefan Dietze
Chapter 9 © Richa Jalota, Nikit Srivastava, Daniel Vollmers, René Speck, Michael Röder, Ricardo Usbeck and Axel-Cyrille Ngonga Ngomo 2020
Chapter 10 © Philips Kokoh Prasetyo, Amila Silva, Ee-Peng Lim and Palakorn Achananuparp 2020
Chapter 11 © Tong Zeng and Daniel Acuna 2020
Chapter 12 © Paco Nathan 2020

First published 2020

This book is published under a Creative Commons Attribution-Non Commercial Share Alike 3.0 Unported (CC BY-NC-SA 3.0) license. You are free to share, copy and redistribute the material in any medium or format and adapt, remix, transform and build upon the material, provided that you give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. You must not use the material for commercial purposes. You must distribute your contributions under the same license as the original if you remix, transform, or build upon the material. The full license can be found at: <https://creativecommons.org/licenses/by-nc-sa/3.0/legalcode>. For further information about reuse, please contact contracts@sagepub.co.uk.

Library of Congress Control Number: 2019920051

British Library Cataloguing in Publication data

A catalogue record for this book is available from the British Library

ISBN 978-1-5297-0586-7 (pbk)

At SAGE we take sustainability seriously. Most of our products are printed in the UK using responsibly sourced papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

CONTENTS

List of Contributors	ix
1 Introduction	1
The Reason for This Book	2
How This Book Came to Be	3
Book Overview	4
General Competition Information	9
Acknowledgements	9
References	9
Notes	10
2 Where's Waldo? Conceptual Issues When Characterizing Data in Empirical Research	11
Introduction	12
Insights from a Research Data Centre Perspective	13
Lessons Learned From the Competition	18

Discussion	23
Conclusions	24
Acknowledgements	25
References	25
Notes	26
3 Digital Science Use Cases: Enriching Context and Enhancing Engagement Around Datasets	27
Introduction	28
Information Infrastructure	29
Cultural Infrastructure	31
Enriching Context	32
References	37
4 Metadata for Social Science Datasets	39
Introduction	40
Data Elements and Datasets	40
Metadata Schemas and Catalogues	41
Linked Data	42
Richer Semantics	42
Data Repositories and Collections of Datasets	43
Repository Services	46
Infrastructure	49
Conclusion	49
Acknowledgements	50
References	50
Notes	51
5 Competition Design	53
Introduction	54
Competition Design	56
Data	57
Submission Process	65
Evaluation	69
Discussion	76
Conclusion	78
Appendix: Standardized Metadata, Full Text and Training/Evaluation for Extraction Models (Alex Wade and Sebastian Kohlmeier)	78
References	81
Notes	82

6 Finding Datasets in Publications: The Allen Institute for Artificial Intelligence Approach	83
Introduction	84
Methods	84
Results	88
Conclusion	90
Acknowledgements	91
References	91
Notes	92
7 Finding Datasets in Publications: The KAIST Approach	93
Non-Technical Overview	94
Literature Review	94
What We Did	94
What Worked and What Did Not	99
Summary of Results and Caveats	100
Lessons Learned and What We Would Do Differently	103
Conclusion	103
What Comes Next	103
Acknowledgements	104
Appendix: Description of the Code and Documentation	104
References	104
8 Knowledge Extraction from Scholarly Publications: The GESIS Contribution to the Rich Context Competition	107
Introduction	108
Approach, Data and Preprocessing	109
Dataset Extraction	111
Research Method Extraction	114
Research Field Classification	120
Discussion and Limitations	122
Conclusion	124
Acknowledgements	124
References	124
Notes	126
9 Finding Datasets in Publications: The University of Paderborn Approach	129
Literature Review	130
Project Architecture	131
Preprocessing	131

Approach	133
Dataset Extraction	135
Evaluation	137
Discussion	138
Future Agenda	139
Appendix	140
References	140
Notes	141
10 Finding Datasets in Publications: The Singapore Management University Approach	143
Introduction	144
Related Work	144
Data Analysis	145
Methods	147
Experiment Results	151
Acknowledgements	154
References	154
Notes	155
Appendix: Technical Documentation	155
11 Finding Datasets in Publications: The Syracuse University Approach	157
Introduction	158
The Dataset	158
The Proposed Method	159
Results	162
Conclusion	163
Acknowledgements	164
References	164
Notes	165
12 The Future of AI in Rich Context	167
The Case For Rich Context	169
Use Cases For Rich Context	171
System Architecture Overview	176
Trends, From Origins To Near-Term Future Projections	178
Summary	178
References	179
Notes	179
Index	181

LIST OF CONTRIBUTORS



Karam Abdulahhad is a postdoctoral researcher 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 PhD degree in computer sciences from Grenoble-Alpes University in France, where he tackled the problem of term mismatch. He proposed a new information retrieval (IR) model by adapting an idea from physics. His research interests include IR theory, logical/conceptual/semantic IR, machine learning, and text mining. He has recently been studying the profitability of modern embedding techniques in IR. He has taught in several universities and developed several tools.



Palakorn Achananuparp is a senior research scientist at the 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, *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 the Bellcore Applied Research group in information science and digital libraries. He was editor-in-chief of *ACM Transactions on Information Systems* and later chair of the ACM Publications Board. Since 1998 he has been a faculty member at a number of universities around the world, including Maryland, Drexel, Victoria (New Zealand), 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 PhD degree in artificial intelligence from Carnegie Mellon University. Before pursuing the PhD, Waleed was a software developer at Microsoft Research, web developer at eSpace Technologies, and teaching assistant at Alexandria University.



Iz Beltagy is a research scientist at the Allen Institute for AI (AI2), working on Natural Language Processing, with a focus on scientific documents. He received his PhD in Computer Science from the University of Texas at Austin in 2016.



Stefan Bender is head of the Research Data and Service Center of the Deutsche Bundesbank and, since 2018, honorary professor at the School of Social Sciences, University of Mannheim. With his position at Deutsche Bundesbank he was chair of INEXDA (the Granular Data Network) and vice-chair of the German Data Forum (www.ratswd.de). Before joining the Deutsche Bundesbank, Stefan was head of

the Research Data Center (RDC) of the Federal Employment Agency at the Institute for Employment Research (IAB), which established an international research data centre, including access to IAB data in the US (for example Berkeley, Harvard). His research interests are data access, data quality, merging administrative, survey data and/or big data, record linkage, management quality and mobility of inventors. He has published over 100 articles, in journals including *American Economic Review* and *The Quarterly Journal of Economics*.



Christine Betts is a software engineer working on human computation at Google AI. She graduated with honours 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 a research associate in the Knowledge Technologies for the Social Sciences department 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 and journalism, and automatic fact-checking.



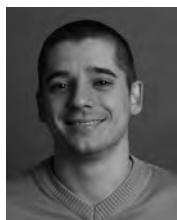
Minh-Son Cao is a master's student in the School of Computing at KAIST, under the supervision of Professor Sung-Hyon Myaeng at the Information Retrieval and Natural Language Processing Laboratory. Previously, he received his bachelor's degree from the University of Engineering and Technology, Vietnam National University (VNU-UET) in June 2017. He was a member of the 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 of data and knowledge engineering at the Institute for Computer Science at Heinrich-Heine-University Düsseldorf, scientific director of the Knowledge Technologies for the Social Sciences department 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 Conference, SIGIR, CHI and ISWC, where he also frequently serves as PC and/or OC member.



Dimitar Dimitrov is a postdoctoral researcher at GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany. He obtained a PhD 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 *dalra* project aimed at delivering 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 behaviour in web-based systems.



Hendrik Christian Doll is a data scientist at the Research Data and Service Centre of Deutsche Bundesbank, where his responsibilities include advancing various data science projects. He is a lecturer for exploratory data analysis at EBS Universität für Wirtschaft und Recht. His professional interests include applying supervised machine learning techniques for record linkage, using text extraction to find data citations in text, and automating data visualization in corporate design. He holds a MSc in Econometrics from the University of Geneva. When not in the office, you will find him rock climbing.



Behnam Ghavimi is a research fellow in the Knowledge Technologies for the Social Sciences 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's thesis, under the supervision of Professor Sören Auer and Dr Philipp Mayr, was about detecting dataset references in texts. Since September 2016, he has been 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 a 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 life cycle for ingest, curation, facilitating data discovery, and providing access to sensitive, administrative data for academics and policy analysts. Andrew has an MS degree in information from the

University of Michigan School of Information and a BA 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 in 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 and Seattle.



Mark Hahnel is the CEO and founder of Figshare, which he created while 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 the potential it has to revolutionize the research community. Since 2012, Mark has led the development of research data infrastructure, with the aim of reusable and interoperable academic

data. Mark sits on the board of DataCite, the advisory board for DOAJ, was a judge for the National Institutes of Health (NIH), Wellcome Trust Open Science prize, and acted as an advisor for the SpringerNature masterclasses.



Christian Herzog is CEO of Dimensions and chief portfolio officer at Digital Science. A medical doctor by training, Christian also studied economics and in 2005 started 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 VP for Product Management SciVal. In 2013, Christian and his co-founders started ÜberResearch as part of Digital

Science, which led to the launch of Dimensions as a large-scale research information infrastructure in 2018.



Christian Hirsch works at the Research Data and Service Centre of Deutsche Bundesbank. Before coming to Bundesbank he was a post-doctoral researcher and head of data center at the research institute Sustainable Architecture for Finance in Europe (SAFE), at Goethe University Frankfurt. He earned his PhD in Finance at Goethe University. Christian's research interests include financial intermediation, monetary policy, and corporate governance. His work has been published in *The Review of Financial Studies* and *Journal of Banking & Finance*.



Daniel W. Hook is CEO of Digital Science. Daniel co-founded Symplectic, a research information management company, in 2003, with his PhD offemates. Symplectic received investment from Digital Science in 2010 and since then Daniel has worked across Digital Science's portfolio. In 2015, he became CEO. Daniel continues to be an active researcher and holds visiting positions in Physics at Imperial College London and Washington University in St Louis, is a Policy Fellow at CSaP at the University of Cambridge, and is Co-Chair of the Research on Research Institute.



Giwon Hong is a master's student in the School of Computing at KAIST and research assistant in the IR&NLP Lab at KAIST. He graduated from Sungkyunkwan University with a degree in computer science in February 2018. His research lies in the area of natural language processing, specifically in question answering and relation extraction.



Richa 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 an applied research scientist on the Semantic Scholar team at the Allen Institute for AI. He previously interned at Microsoft and Facebook, and received his BS degree in computer science from Harvey Mudd College in May 2018. His research interests are in natural language processing and using AI techniques to make useful tools for humans.



Sebastian Kohlmeier is the senior manager of programme management and business operations at the Allen Institute for AI, where he leads programme 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 programme manager and engineering manager in a variety of roles at Amazon and Microsoft. Sebastian graduated with honours from Western Washington University in 2007.



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



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.



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 co-founded the Coleridge Initiative, whose goal is to use data to transform the way governments access and employ data for the social good through training programmes, research projects and a secure data facility. The approach is attracting attention throughout the USA, 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 social science research, with a 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 in projects with social impact and utilizing text as data in a variety of applications for social science research.



Ee-Peng Lim is the Lee Kong Chian professor of information systems and director of the Living Analytics Research Center at 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, many of them in top ACM and IEEE journals, and presented 280 conference papers. He received 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 BA in computer science from Wittenberg University, an MA in journalism from NYU, and was a University Enrichment Fellow at Michigan State University.



Ian Mulvany is CTO at BMJ, where he leads teams that deliver the technologies and features that underpin new products as well as underlying publishing systems. He is responsible for leading on technology strategy for the company and understanding the implications of new and emerging technologies for the business. Previously, he helped set-up SAGE's methods innovation incubator SAGE Ocean following a lean product development approach. Prior to SAGE 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.



Sung-Hyon Myaeng is a professor of Computer Science in the School of Computing and the head of Web Science and Technology Division at Korea Advanced Institute of Science and Technology (KAIST). He is also the Director of KAIST-Microsoft Research Collaboration Center (KMCC). Previously he was on the faculty at Syracuse University, USA, where he was granted tenure in 1994. He earned his MS and

PhD from Southern Methodist University, Texas, USA in 1985 and 1987, respectively. His research has been in the intersection between lexical and semantic aspects in natural language processing and unconventional search techniques in information retrieval.



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 in the Computer Science department of Paderborn University. In his work he focuses on the life cycle of knowledge graphs. He has been involved in the development of approaches for the extraction, storage, querying, integration, fusion and 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. Since finishing with a master's degree at the NLP Group at Leipzig University (Prof. Dr. Gerhard Heyer), he has been part of a team in a third-party funded project (German Research Fund) to build up a specialized information service for political scientists (pollux-fid.de), a project of the State and University Library Bremen (SuUB) in cooperation with GESIS. During his studies, he collaborated on 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 programmes; and in the Division of Disease Control, working with real-time emergency department, reportable infectious disease, and school health data. Sophie holds a Bachelor of Science in Engineering degree 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 PhD 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 various scientific journals and conferences.



Haritz Puerto San Roman is a master's student in the School of Computing at KAIST and research assistant in the IR&NLP Lab at KAIST. He graduated from the University of Malaga with a degree in computer science in July 2017. His research lies in the application of machine learning to natural language processing, specifically to solve the problem of question answering.



Amila Silva graduated from the University of Moratuwa, Sri Lanka, with a first-class honours degree in electronics and telecommunication engineering, where he was placed second of the graduating class of 110 students. He is currently working towards a PhD degree at the Department of Computing and Information Systems, University of Melbourne, Australia. He was awarded the Melbourne Graduate Research Scholarship to support his studies. He was also awarded the Rowden White Scholarship, a prestigious scholarship provided by the University of Melbourne to talented PhD students. His research interests include continual learning, graph analytics, and data mining.



René Speck is a research associate and a PhD 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 Leipzig University since 2013. He has also since then been a reviewer for several conferences and journals.



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 is involved 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 in 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 PhD 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 programme manager for knowledge graphs and open science at the Chan Zuckerberg Initiative. Alex earned his master's degree 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.



Duane E. Williams is a vice president of US Government at Digital Science. Duane earned his doctorate in theoretical chemistry from the Quantum Theory Project at the University of Florida. His work focuses on improving strategic research investment decisions by using new data sets, tools and metrics to gain greater insight into the global research landscape. Prior to joining Digital Science, he served as a 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.



Tong Zeng is a PhD 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 has also served as a lecturer on 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.



Madeleine van Zuylen is a data science analyst at the Allen Institute for AI (AI2) on the Semantic Scholar team. Before joining AI2, Madeleine graduated from the University of Notre Dame in 2017 with a BS in Biochemistry and Applied Computational Mathematics and Statistics.

1

INTRODUCTION

Ian Mulvany, Paco Nathan,
Sophie Rand, and Julia Lane

The Reason for This Book	2
How This Book Came to Be	3
Book Overview	4
General Competition Information	9
Acknowledgements	9
References	9
Notes	10

The Reason for This Book

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 (Ioannidis, 2005; Munafò et al., 2017). There is an enormous interest in changing 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 (Colavizza et al., 2019).

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. (Bush, 1945)

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

A major 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 is being used, in what research fields, with what methods, with what code and with what findings – often by passively capitalizing on the accumulated labour 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 (Yarkoni et al., 2019).

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 (Ioannidis, 2005; Young et al., 2008; Dafoe, 2014; Camerer et al., 2018; Christensen and Miguel, 2018) is even more acute in the study of human behaviour, 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 and data storage to statistical analysis and dissemination of results (Yarkoni et al., 2019). And, in the United States, the recent passage of the Foundations of Evidence-Based Policymaking Act (Hart & Shaw, 2018; Office of Management and Budget, 2019) and the focus on a Federal Data Strategy mean that there is an important use case for showcasing the value of new approaches.

Knowing how data have been produced and used before is important for the reproducibility and replicability of empirical research. Unfortunately, it has been difficult to get information about the key elements, namely, what does the data *measure*, what *research* has been done by what *researchers*, with what *code*, and with what *results*. The acquisition of 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.

This book is about creating a new, automated approach to collecting such rich contextual metadata around datasets – what we call ‘Rich Context’.

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 Evidence-Based Policymaking Commission Act that established a Commission to make a set of recommendations on how better to use data for decision-making (https://en.wikipedia.org/wiki/U.S._Commission_on_Evidence-Based_Policymaking). 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 do to advance the field, and received support from Schmidt Futures, the Alfred P. Sloan Foundation and the Overdeck Family Foundation

for support. As part of that support, we ran what we called a ‘Rich Context’ 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 goal of the competition 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 15 February 2019.¹

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 knowledge graph of moderate size. 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.

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. The other chapters in this book thus 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 I provides an overview of the motivation and approach. Section II describes new approaches to develop corpora and ontologies. Section III describes the competition results in terms of model development. Section IV provides a forward-looking agenda.

Section I: Motivation and Approach

In Chapter 2, researchers from the Research Data and Service Center at the Deutsche Bundesbank show us why better metadata for social science data enables the 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 meta-data: 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 are 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 user's perspective helps the data creator to improve upon the quality of the data itself, improving dataset owners' service delivery (e.g. the 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 the 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 also the interplay between them – 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.

In Chapter 3 researchers from Digital Science describe the role user engagement plays in creating rich context around datasets, which take on properties of 'first-class research objects' (such as journal articles) in that they are published as distinct research outputs in their own right. The authors lay out a set of challenges in the sharing of datasets and dissemination of dataset metadata, and articulate goals in creating an infrastructure to answer these challenges. They argue for the creation of two aspects: 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 analyse 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 meta-data on datasets. Cultural infrastructure around datasets does not fit into the existent culture of research publications.

Section II: New Approaches

A major challenge faced by this new field is 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 4, 'Metadata for Social Science

Datasets', Robert B. Allen describes a framework for the application of metadata to datasets, details existing metadata schemas, 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, such as 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, such as search, computing environments, preservation of data for archives, and logging of the history of a dataset and its provenance. Allen describes other challenges 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, such as the number of downloads and maintenance of datasets.

Chapter 5 describes the details of the Rich Context Competition. 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.

The chapter describes the implementation of the competition and provides a number of pointers to useful open source code and tools that were developed by the authors. It explains the development of the 'gold standard corpus' (GSC) of academic publications tagged with the semantic context within which datasets are mentioned used in analysis. It also describes the design of the competition, in which competitors were provided with two sets of data related to detecting datasets. The first was a catalogue of all of the datasets 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. The second was a subset of these datasets that were actually specifically annotated as having been used for analysis in a given publication. The chapter also provides links to the content analysis protocol and accompanying web-based coding application that was developed to enable human coders to examine all of the dataset citations in our train-test corpus and identify dataset mentions. We also challenged the teams to detect methods and fields for a given publication that not only can determine when a given publication is a part of an existing field or uses an existing method, but could also detect new fields and methods as they emerge.

The chapter also reviews the technology that was developed to enable participants to build and submit a model. It explains in some detail the options that were available and the reasons for the choice of a combination of Box.com, Docker, a Git repository for code to implement and support infrastructure, and shell scripts. It also describes the submission framework which used a file-system-based API for giving the model input and accepting output and allowed the team to interact through the file system to keep the configuration and implementation simple. Another contribution of the chapter is that it provides details about how the team was able to meet the goal of having 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. There is a discussion of the choice of a container-based approach which simplified the running of models, and allowed for flexibility and configurability.

Finally, the chapter describes the evaluation approach, in terms of both the judges and the development of a coder tool that was used to make judgment decisions. It concludes with lessons learned.

The appendix to the chapter, by Alex Wade and Sebastian Kohlmeier, provides even more detail on how to improve future competitions. The Rich Context Competition made use of ad-hoc processes to identify metadata and full text sources for their specific tasks, which resulted in inconsistencies and a lack of versioning of input data across competitions and projects. They provide an overview of how these challenges can be reduced by using a trustworthy metadata source such as Semantic Scholar's open corpus² developed by the Allen Institute for Artificial Intelligence (AI2), or Microsoft's Academic Graph³ that makes it easy to access standardized metadata from an openly accessible source.

Section III: Competition Results

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 tasks. 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 in identifying research fields, by using Wikipedia as the source, and methods by using machine learning techniques.

Chapter 8, by the GESIS team, also used a named entity recognition procedure. However, their design followed a module-based approach and they developed tools that can be used separately but also as parts of a data processing pipeline. To identify 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 an implicit entity linking approach to tackle dataset extraction tasks. They employ weakly supervised classification for research method and field identification tasks utilizing Sage Knowledge as an external source and as a proxy for weak labels.

Chapter 11, by Syracuse University researchers, describes a method for extracting dataset mentions based on a deep recurrent neural network. The team used a bidirectional long short-term memory sequence to sequence model paired with a conditional random field (CRF) inference mechanism. The architecture is similar to that of Chapter 6, but focussed only on the detection of dataset mentions. The model achieved a relatively good performance, but there were a number of limitations which are carefully discussed in the chapter.

Section IV: Looking Forward

The future agenda is described in the concluding chapter by Paco Nathan. 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 catalogue. 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 to the system. With the latter step, in the broader sense rich context helps establish a community focused on contributing code and knowledge to the research process.

General Competition Information

The competition had two phases. In the first phase, participants were provided with labelled data, consisting of a corpus of 2500 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 unlabelled 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:

- 30 September 2018. Participants submit a letter of intent.⁴
- 15 October 2018. Participants notified and Phase 1 data provided.⁵
- 15 November 2018. Preliminary algorithm submitted.⁶
- 1 December 2018. Fifteen finalists selected⁷ and Phase 2 data provided.⁸
- 15 January 2019. The algorithms of up to six teams are selected for final submission.⁹
- 15 February 2019. Workshop is held in New York for final presentation and selection of winning algorithms.¹⁰

All the information provided to participants was available on GitHub.¹¹

Acknowledgements

The support of Schmidt Futures, the Alfred P. Sloan Foundation and Overdeck Family Foundation is gratefully acknowledged.

References

- Bush, V. (1945) *Science, the Endless Frontier: A Report to the President*. Washington, DC: US Government Printing Office.
- Camerer, C. F. et al. (2018) Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, **2**, 637.
- Christensen, G. and Miguel, E. (2018) Transparency, reproducibility, and the credibility of economics research. *J. Econ. Lit.*, **56**, 920–980.
- Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. and McGillivray, B. (2019) The citation advantage of linking publications to research data. Preprint, arXiv:1907.02565.
- Dafoe, A (2014) Science deserves better: The imperative to share complete replication files. *PS Polit. Sci. Polit.*, **47**, 60–66.

- Hart, N. and Shaw, T. (2018) Congress provides new foundation for evidence-based policymaking (available at <https://bipartisanpolicy.org/blog/congress-provides-new-foundation-for-evidence-based-policy-making/>).
- Ioannidis, J. P. A. (2005) Why most published research findings are false. *PLoS Medicine*, **2**, e124.
- Munafò, M. R. et al. (2017) A manifesto for reproducible science. *Nature Human Behaviour*, **1**, 21.
- Office of Management and Budget (2019) Federal Data Strategy (available at <https://strategy.data.gov>).
- Wissler, L., Almashraee, M., Díaz, D. M. and Paschke, A. (2014) in *IEEE GSC*.
- Yarkoni, T. et al. (2019) 'Enhancing and accelerating social science via automation: Challenges and opportunities', Preprint, SocArXiv, doi: 10.31235/osf.io/vncwe.
- Young, N. S., Ioannidis, J. P. A. and Al-Ubaydli, O. (2008) Why current publication practices may distort science. *PLoS Medicine*, **5**(10), e201.

Notes

- 1 <https://coleridgeinitiative.org/richcontextcompetition/workshopagenda>
- 2 <https://www.semanticscholar.org>
- 3 <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/>
- 4 <https://coleridgeinitiative.org/richcontextcompetition#howtoparticipate>
- 5 <https://coleridgeinitiative.org/richcontextcompetition#phase1participation>
- 6 <https://coleridgeinitiative.org/richcontextcompetition#programreqs>
- 7 <https://coleridgeinitiative.org/richcontextcompetition#phase1evaluation>
- 8 <https://coleridgeinitiative.org/richcontextcompetition#phase2participation>
- 9 <https://coleridgeinitiative.org/richcontextcompetition#phase2evaluation>
- 10 <https://coleridgeinitiative.org/richcontextcompetition#phase2evaluation>
- 11 <https://github.com/Coleridge-Initiative/rich-context-competition>

2

WHERE'S WALDO? CONCEPTUAL ISSUES WHEN CHARACTERIZING DATA IN EMPIRICAL RESEARCH

Stefan Bender, Hendrik Christian Doll and
Christian Hirsch

Introduction.....	12
Insights from a Research Data Centre Perspective.....	13
Lessons Learned From the Competition	18
Discussion.....	23
Conclusions.....	24
Acknowledgements.....	25
References.....	25
Notes	26

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 developments. 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 the user experience. One specific project currently pursued 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 distinguished 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 labour-intensive and error-prone. Thus, the competition to automate the discovery of datasets, the results of which are described in this book, together with associated research methods and fields in social science research publications, lays the groundwork for any future implementation of such a recommendation system.

In this chapter we present lessons learned from the competition described in this book. We do this from our collective background in social science. Readers interested in the more technical aspects of the competition may find Chapters 6–11 helpful, in which 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 from designing such a competition. Finally, Chapter 12 presents in more detail a potential framework for implementing a microdata recommendation system.

Extracting dataset citations from publications is a fairly difficult task because of the variety of dataset citation formats and the absence of training data. Besides empirical research support, the information gained is the basis for providing value for policy purposes in the G20 context. For example, by providing researchers with information about the use and availability of microdata 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 digitization 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 present 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

Research data centres represent an established operational approach to facilitate access to confidential microdata for statistical purposes. This approach is based on the Five Safes theoretical framework which was initially developed by Felix Ritchie at the UK Office of National Statistics in 2003 (see Desai et al., 2016). The first dimension refers to safe projects. This dimension mainly refers to whether the intended use of the data conforms with the use specified in legislation or regulations. For example, legislation may specifically allow users to use the data only for independent scientific research.

Safe people, the second dimension of the Five Safes 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 level may be required to access the data. In contrast, safe data, the third dimension, refers to the potential to de-identify individuals or entities in the data. Safe settings, the fourth dimension, relate to the practical controls on how the data is accessed. Different channels may exist which in turn may depend on the de-identification risk. In practice, the lower the de-identification risk the more restrictive the setting will be. Safe output, the fifth and final dimension, refers to the risk of de-identification in publications from confidential microdata.

In response to the increased internal and external demand for microdata and to data confidentiality requirements, in 2013 the Bundesbank set up the Integrated Microdata-based Information and Analysis System (IMIDIAS) and established the RDSC (for more detail, refer to Kalckreuth, 2014; Bender and Staab, 2015). Many of the technical and organizational measures put in place at the RDSC to protect confidential microdata follow the Five Safes framework.

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

The data access provided by the RDSC and the underlying legal requirements are described in detail by Schönberg (2018). In a nutshell, users' requests to use microdata are first reviewed in line with legal requirements. Once their application 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 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. Alternatively, for most of the Bundesbank's surveys – such as the Panel on Household Finances 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 is documented in detail and archived. In doing so, the RDSC works according to globally recognized standards and was accredited as a research data centre by the German Data Forum (Rat für Sozial- und Wirtschaftsdaten).

The Knowledge-Generating Process of Empirical Research at the RDSC

In this section we present the knowledge-generating process of empirical research at the RDSC. This knowledge-generating process can be organized along four key dimensions: data services, research, publications, and (structured) user-specific knowledge (Figure 2.1). For the sake of simplicity, we assume that data services affect research and that the outcome of research is a publication. From a publication the RDSC (or the Bundesbank) is able to distil 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 may elevate the knowledge-generating process to a higher level.

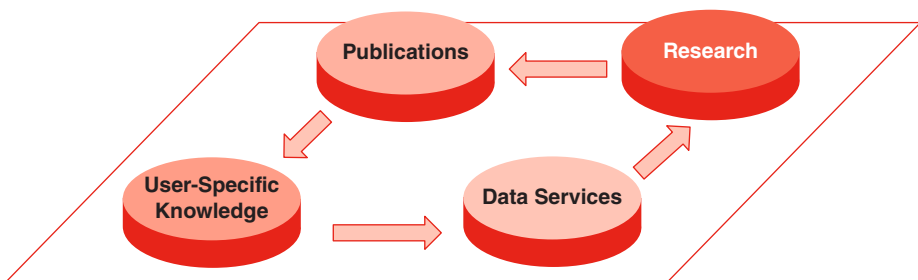


Figure 2.1 The four dimensions of the knowledge-generating process of empirical research at the RDSC

The knowledge-generating process starts with the data services offered by the RDSC to researchers, who are analysing data from the RDSC. Data services comprise raw microdata and comprehensive documentation of the data, both of which the RDSC compiles together with the data-producing units at the Bundesbank.¹ Furthermore, the data services dimension also includes the methodological improvement of microdata through, for example, 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 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), for example, 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 ultimately used
- What information about data is most valuable to get to the results, or what kind of 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 data services to be improved by making discovery of data and related projects, people, and publications at the Bundesbank more comprehensive and efficient. For example, knowledge harvested from publications may be used to enhance services provided by the RDSC by allowing standard datasets to be tailored to the needs of researchers. 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 2.2. 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.

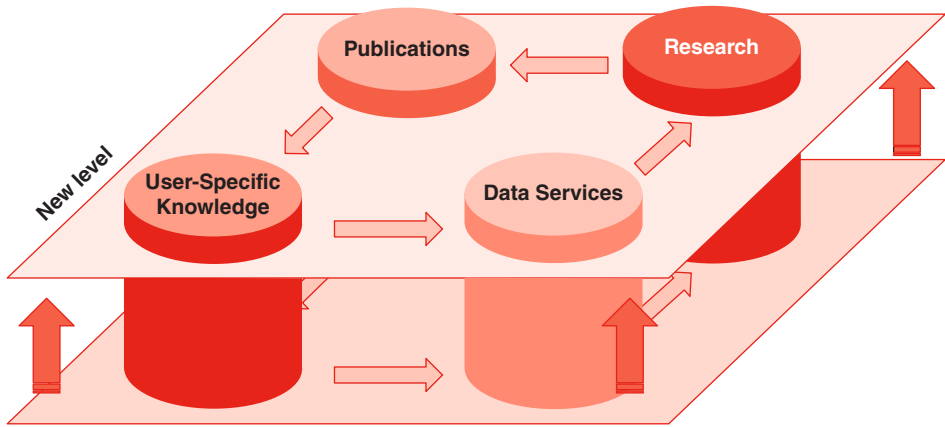


Figure 2.2 Elevating the knowledge-generating process of empirical research at the RDSC to a higher level by enabling a feedback loop to data services

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 knowledge available in publications to be transformed into generally reusable knowledge to inform stakeholders (data producers, RDSC, Bundesbank decision-makers). 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 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 strengthened. Positive externalities (i.e. research as a public good) suggests a less than 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 demonstrating a dataset's impact.

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

Furthermore, the usability index can be expanded into a measure of 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 such as ‘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 such as ‘users like you also bought D’.

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

However, we suspect a simpler underlying behavioural 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, fewer 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 4 of this book.

A challenge in building a data-driven recommendation system is to make sure that recommended datasets are indeed feasible to use, that is, constitute meaningful recommendations. Thus, besides information about datasets, additional information, such as fields and methods, needs to be ingested into the system. This additional information essentially constitutes additional links between datasets that help 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 researchers data access free of charge should be backed by empirically measurable benefits 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 for approximating the societal benefits of data access can be to measure the creation of knowledge² resulting from specific datasets.

One can argue that the 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 granular dataset which enables causal evidence

adds more value to societal knowledge, than a comparable dataset that does not allow identification of causal effects. However, both of these methods require identifying which empirical result from a publication can be attributed to which dataset.

Lessons Learned From the 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 employ 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 strings and following entity resolution (matching the string to the correct entity or digital object identifier (DOI)). Concerning entity resolution, they report the difficulty of broad survey dataset citations that ignore data variability (such as years, versions, questionnaire variants), motivating a dataset taxonomy. Named dataset citations are often underspecified, allowing identification of the survey but not of the precise dataset (which of multiple subsamples, aggregation levels, survey modes, etc.).

Zhang et al. (2016) use a bootstrapping approach to extract dataset citations from 116 computer science 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 a formatting proposal for standardized data citations; however, their categories do not optimally support automatic search and extraction, if no unique dataset identification (such as a DOI) is used.³ In the context 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 provided by the RDSC are all DOI registered.

Improving dataset citation is high on the scientific agenda in recent years. This notably includes promoting 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 during the competition described in this book. These lessons originated from our motivation presented in the previous section 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.⁴

We organize this section around the three sets of information that were the main focus of the competition: dataset 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. 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 have a well-defined scope, thus allowing clear citation, usually including a unique dataset identifier (such as a DOI).

Lesson 1: Datasets Fall into Two Broad Categories

Since the corpus of publications used for the competition spanned different domains (health care, education, and others), we quickly realized that our dataset image had an econocentric bias. In social science, we learned, datasets can be divided into two broad categories for the purposes of extraction.

First, there are named datasets, that is to say, well-defined, usually large-scale and publicized datasets (e.g. Compustat). Generally, named dataset mentions are short strings in the publications, have commonly used abbreviations (e.g. MMSR), and often contain the name of an institution or commercial data vendor. Sometimes (rarely, but increasingly) these datasets can be identified by a unique 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 used across multiple papers and research domains.

The second dataset category is what we call ‘created datasets’. These are datasets usually collected or built by authors of a publication for the purpose of analysing one specific

research 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 a brief description (of one or more paragraphs) 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 such as ‘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 commonly used abbreviation. Data collection is usually paper specific, and the universe of existing datasets is 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 data sources predominantly used naturally vary. This has important repercussions for designing a competition, since algorithm performance and later recommendation system performance vary 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 whether datasets are named or created. 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 fields 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 this 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 the 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 different time periods or subsamples, different states of data and states of knowledge, computational data preprocessing or enrichment steps, to name but a few. These identification issues complicate the current task of entity resolution of extracted dataset mentions.

Unique dataset identification carries significant repercussions for reproducibility purposes, where identifying 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 reproducible. To make lives easier, DOIs also greatly facilitate the automated extraction of well-defined datasets from publications (comparable to largely standardized citations of other publications, allowing easy retrieval of publication networks, etc.). An alternative approach to ensure identification of datasets could be to provide richer, more systematic metadata for datasets. See Chapter 4 in this book for a detailed discussion of this point.

To summarize, 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 of the use of DOIs is harder in this case, because of a larger target group – authors rather than a limited number of data stewards. Even in the case of widespread DOI usage for named datasets, the competition algorithms yield valuable results through the extraction of created datasets in order to allow datasets used in the past to be referenced and made available for further analysis.

Lesson 4: Datasets Mentions Could Indicate Datasets Used for Analysis or 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; for example, 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 the existence of publication networks and publications maps), 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 them ('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, and so on in the corpus of publications.⁵ Obviously, this is helpful for algorithms to correctly identify true negatives, that is, 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 depends on such a distinction.

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 deliberately did not provide any thesauri to the competition. 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). On the other hand, thesauri limit the catalogue of potentially identifiable fields and methods, thus prohibiting new methods and fields from being identified in fast-changing modern research areas. Also thesauri might disturb algorithm performance, since algorithms might be forced to place topics and fields in 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 uninformative) and too narrow predictions (narrow, but potentially wrong). A potential way out of this is backward induction: 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 (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 decided to include in statistical methods all methods that describe the potential for a merge of datasets or joint usability, hence to include all those listed above. We considered a broad definition of methods, including not only high-level statistical methods, such as ordinary least squares, but also the observed unit, time period or even regression 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 insights.

Discussion

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

We would have had 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 are only now and slowly gaining widespread application for datasets in social science, our task is a one-to- 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 between named and 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 and created datasets varies across research areas.

Also varying across research areas is the number of datasets used per empirical paper. This number also 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 fields 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 investment 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; the circle thus 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.⁶

Automatic processing of generated knowledge in publications is becoming increasingly available with modern text analysis tools. Extracting such information is important,

because timely and optimal usage of gained results increases the speed with which findings can be incorporated into data services, thereby in turn enabling next-level research. To further improve automatic processing, minimum standards for dataset taxonomy are needed. Harmonized metadata schemas for datasets – 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 prototype algorithms delivered 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. It is easier to implement prototyping dataset recommendation systems, usability indices, and so on for research areas with a largely formalized dataset citation and sharing culture, such as natural sciences. However, ultimately the potential for benefits may well be larger in social sciences, where there is much left to be gained.

Acknowledgements

We would like to thank Jannick Blaschke, Rafael Beier, Julia Lane, Paco Nathan, and Ian Mulvany for helpful comments on previous drafts of this chapter. We would also 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.

References

- Ball, A. and Duke, M. (2011) *How to Cite Datasets and Link to Publications*. Edinburgh: Digital Curation Centre.
- Bender, S., Hausstein, B. and Hirsch, C. (2018) An Introduction to INEXDA's Metadata Schema. Technical Report 2018-02, Deutsche Bundesbank, Research Data and Service Centre.
- Bender, S. and P. Staab, P. (2015) The Bundesbank's Research Data and Service Center (RDSC), Gateway to treasures of microdata on the German financial system. IFC Bulletin No 41, Irving-Fisher Committee on Central Bank Statistics.
- Boland, K. and B. Mathiak, N. (2015). Challenges in matching dataset citation strings to datasets in social science. *D-Lib Magazine*, 21, 1/2.

- Boland, K., Ritze, D., Eckert, K. and Mathiak, B. (2012) Identifying references to datasets in publications. In P. Zaphiris, G. Buchanan, E. Rasmussen and F. Loizides (eds), *Theory and Practice of Digital Libraries*. Berlin: Springer, pp. 150–161.
- Desai, T. R., Ritchie, F. and Welpton, R. (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. and Lange, C. (2016). Identifying and improving dataset references in social sciences full texts. Preprint, arXiv:1603.01774.
- Helbig, K., Hausstein, B., Koch, U., Meichsner, J. and Kempf, A. (2014) da|ra Metadata Schema. GESIS Technical Reports 2014-17. doi: 10.4232/10.mdsdoc.3.1
- Kalckreuth, U. von (2014) A Research Data and Service Centre (RDSC) at the Deutsche Bundesbank – a draft concept. IFC Bulletin No 37, Irving-Fisher Committee on Central Bank Statistics.
- Koesten, L., Mayr, P., Groth, P., Simperl, E. and M. de Rijke, M. (2019) Report on the DATA: SEARCH'18 workshop-Searching Data on the Web. *ACM SIGIR Forum*, 52(1), 117–124.
- McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N. and 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. and Newton, M. P. (2012) The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication*, 1(1), eP1035.
- Schönberg, T. (2018) Data access to micro data of the Deutsche Bundesbank. Bundesbank Technical Report 2018–01.
- Lagoze, C. and L. Vilhuber (2017). O Privacy, Where Art Thou? Making Confidential Data Part of Reproducible Research. *Chance*, 30(3), 68–72.
- Zhang, Q., Cheng, Q., Huang, Y. and 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.

Notes

- 1 Data producers in different departments across Bundesbank compile data, (e.g. micro-data, indicators or time series).
- 2 In our model, we have called this knowledge ‘user-specific knowledge’. Here the knowledge is specific in the sense that it can be used to better fulfil the tasks of the Bundesbank.
- 3 Chapter 4 discusses in more detail how metadata may support the discovery of microdata.
- 4 See Chapter 5 for a more detailed description of the evaluation process used in the competition.
- 5 Chapter 12 discusses potential strategies for collecting training data in more detail.
- 6 For the future, ongoing effort is needed to support all four pillars of the level model. The current competition strengthens the connection from the publication pillar to the structured knowledge pillar. In turn, data services can be improved. To support the data services pillar, digital research data centre environments with facilitated access processes (e.g. a ‘data stewardship module’) will in our view improve data access.

3

DIGITAL SCIENCE USE CASES: ENRICHING CONTEXT AND ENHANCING ENGAGEMENT AROUND DATASETS

Christian Herzog, Daniel W. Hook, Mark
Hahnel, Stacy Konkiel and Duane E. Williams

Introduction.....	28
Information Infrastructure	29
Cultural Infrastructure.....	31
Enriching Context.....	32
References.....	37

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. The first is 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. The second is 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 marginalized 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 and their data.

In the hard sciences, the CERN Open Data Portal contains 131 datasets describing particle collisions, each of which comprises 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 datasets (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, these examples do not begin to quantify the amount of detailed personal data available to companies such as Facebook. Data of the 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 new sources.

Many researchers now work with large volumes of data. Fortunately, many facilitating technologies have become commoditized 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 recognizing not only that science is at a crossroads but also 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 not only enrich information infrastructure around datasets, but 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 the incentives, triggers and frameworks that encourage the dataset stewards, experts and users to contribute to these critical information elements.

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 to expect articles to be grouped into journals and published on a specific date, and we expect there to be 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 normalized research publication. Research fields that rely on data are beginning to publish data as an output distinct 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. Visualization 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-supported 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 was 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 (e.g. 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, including: weakness of a homogeneous metadata infrastructure for datasets; inhomogeneity in the types of data that can be shared; proliferation of different platforms to store data; lack of standardized publication practices; and 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 contextualized 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.

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 to 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; and 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 undervalued 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 recognizing 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 incentivized 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 practised, 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 is often not shared over concerns for a perceived legal barrier.

Other concerns are ethical – for example, should this 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 anonymized as it may first appear (Siddle, 2014). Anonymization 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 reuse 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 is not shared due to the information that can be inferred by processing it.

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 experts in data handling. The work associated with making a dataset human-understandable, reproducible and fully contextualized 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 prioritize 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.

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 called 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 recognized 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 contextualized needs to be introduced. In this section, we focus on this second challenge.

The ability to add context to 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 contextualization 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), 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 Figure 3.1).

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).

The Colors of the Donut

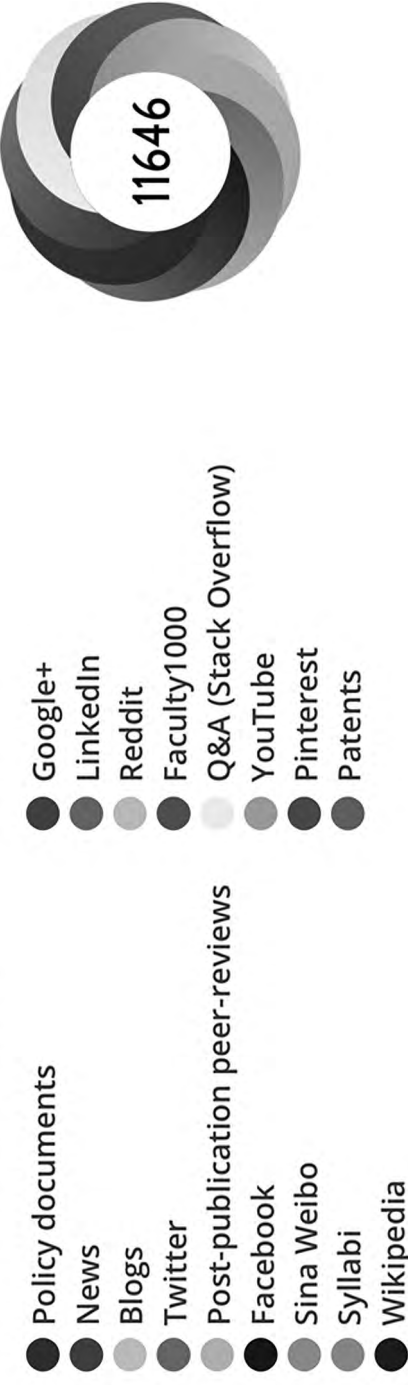


Figure 3.1 Different types of context tracked by Altmetric.com for any research output (reproduced by permission of Altmetric.com)

Correctly developed and accepted by the community, this type of information can make a contribution to solving many of the problems highlighted in this chapter. 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 analysed (e.g. Gigantum, GitHub), we can start to integrate these other crucial engagement contexts as well.

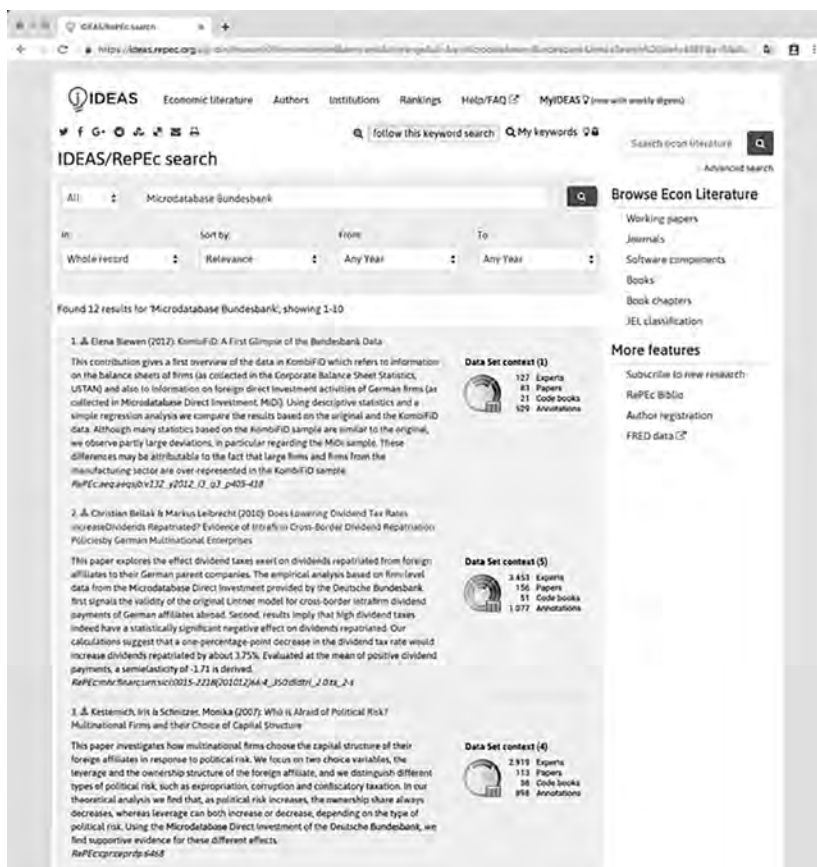


Figure 3.2 Mock-up of a research data badge helping to contextualize a set of search results (Google and the Google logo are registered trademarks of Google LLC, used with permission)



Figure 3.3 Mock-up of a research data badge helping to contextualize a specific dataset

In Figures 3.2 and 3.3 we have visualized some early concepts for how a contextualized research data badge could look. This visualization 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 white paper or policy document
- *Academics who 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. and Altman, M. (2014) Publishing: Credit where credit is due. *Nature*, 508, 312–313. <https://doi.org/10.1038/508312a>
- CERN (2019) CERN Open Data Portal. <http://opendata.cern.ch/search?page=1&size=20&subtype=Collision&type=Dataset> (accessed 30 November 2019).
- CISER, (2019) CISER. URL <https://ciser.cornell.edu/data/data-archive/> (accessed 30 November 2019).
- DataCite Metadata Working Group (2016) DataCite Metadata Schema 4.0. <https://support.datacite.org/docs/schema-40> (accessed 30 November 2019).
- Eijnatten, J. van, Pieters, T. and Verheul, J. (2013) Big data for global history: The transformative promise of digital humanities. *BMGN – Low Countries Historical Review*, 128(4), 5577. <https://doi.org/10.18352/bmgn-lchr.9350>
- Giuliano, F. (2019) Humanités numériques et archives: la longue émergence d'un nouveau paradigme. *Documentation et Bibliothèques*, 65(2), 37–46. <https://doi.org/10.7202/1063788ar>
- Grusin, R. (2014) The dark side of digital humanities: Dispatches from two recent MLA conventions. *differences*, 25, 79–92. <https://doi.org/10.1215/10407391-2420009>
- Hook, D. W., Porter, S. J. and Herzog, C. (2018) Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3, 23. <https://doi.org/10.3389/frma.2018.00023>
- Jisc (2019) Research Data Management Toolkit. <https://rdmtoolkit.jisc.ac.uk/plan-and-design/data-management-planning/> (accessed 30 November 2019).
- Jordan, K. and 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), 1. <https://doi.org/10.5334/jime.448>
- Kamp, K., Herbell, K., Magginis, W.H., Berry, D. and Given, B. (2019) Facebook recruitment and the protection of human subjects. *Western Journal of Nursing Research*, 41, 1270–1281. <https://doi.org/10.1177/0193945919828108>
- Leurs, K. and Shepherd, T. (2017) 15. Datafication & discrimination. In Mirko Tobias Schäfer and Karin van Es (eds), *The Datafied Society: Studying Culture through Data*. Amsterdam: Amsterdam University Press. <https://doi.org/10.1515/9789048531011-018>
- Li, N., Li, T. and Venkatasubramanian, S. (2007) *t*-Closeness: Privacy Beyond *k*-Anonymity and *I*-Diversity. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, pp. 106–115. <https://doi.org/10.1109/ICDE.2007.367856>

- Linek, S. B., Fecher, B., Friesike, S. and 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.0183216>
- NNLM (2019) Data Management Plan | NNLM [WWW Document]. URL <https://nnlm.gov/data/data-management-plan> (accessed 30 November 2019).
- Siddle, J. (2014) I know where you were last summer: London's public bike data is telling everyone where you've been. <https://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html> (accessed 30 November 2019).
- Stark, L. (2018) Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, 48, 204–231. <https://doi.org/10.1177/0306312718772094>
- Treadway, J., Hahnel, M., Leonelli, S., Penny, D., Groenewegen, D., Miyairi, N., Hayashi, K., O'Donnell, D., and Hook, D., 2016. The State of Open Data Report. *Digital Science*. <https://doi.org/10.6084/M9.FIGSHARE.4036398.V1>

4

METADATA FOR SOCIAL SCIENCE DATASETS

Robert B. Allen

Introduction.....	40
Data Elements and Datasets	40
Metadata Schemas and Catalogues.....	41
Linked Data	42
Richer Semantics.....	42
Data Repositories and Collections of Datasets	43
Repository Services	46
Infrastructure	49
Conclusion.....	49
Acknowledgements.....	50
References.....	50
Notes	51

Introduction

Evidence-based policy needs relevant data (Commission on Evidence-Based Policy-making, 2017; Lane, 2016). Such data is often difficult to find and use. The FAIR open access guidelines suggest that data should ideally 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, and then the basic principles of metadata and their application to datasets. Modern metadata is often implemented with Resource Description Framework (RDF) linked data. Next, we introduce ontologies and other semantic approaches. We then move on to applications which use metadata. We examine repositories that hold and distribute collections of datasets. We then describe services and techniques associated with repositories. We conclude by briefly describing 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)² as ‘a collection of data, published or curated by a single agent’³ 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,⁴ concise structured descriptions can be developed for how data elements are combined to form datasets.

Metadata Schemas and Catalogues

Many datasets are available; the DataCite repository alone contains over 5 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.⁵ 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 ORCID⁶) identifies the digital object or its creator.⁷ Metadata elements are generally part of a schema or frame. DCAT⁸ is a schema standard for datasets that is used by many repositories such as data.gov. Other structured frameworks for datasets include the DataCite⁹ metadata schema and the Inter-university Consortium for Political and Social Research Data Documentation Initiative (ICPSR DDI; see below). ISO 19115–1:2014 establishes a schema for describing geographic information and services.¹⁰

The schema specifications provide a flexible framework. For instance, DCAT allows the inclusion of metadata elements drawn from domain schemas and ontologies. Some of these domain schemas are widely used resources which DCAT refers to as ‘assets’. Figure 4.1 shows a fragment of properties (i.e. metadata elements) from an implementation of the Schema.org¹¹ dataset schema to describe gross domestic product (GDP).

```
{
  "name": "gdp",
  "title": "Country, Regional and World GDP (Gross Domestic Product)",
  "description": "Country, regional and world GDP in current US Dollars ($). Regional means collections of countries e.g. Europe & Central Asia. Data is sourced from the World Bank and turned into a standard normalized CSV.",
  "image": "http://assets.okfn.org/p/opendatahandbook/img/data-wrench.png",
  "readme": "Country, regional and world GDP in current US Dollars ($). Regional means collections of countries e.g. Europe & Central Asia.\n\n## Data\n\nThe data is sourced from the World Bank (specifically [this dataset][current]) which in turn lists as sources: *World Bank national accounts data, and OECD National Accounts data files*. \n\nNote that there are a variety of different GDP indicators on offer from the World Bank including:\n\n* [GDP in current USD][current]\n\n* [GDP in constant USD (2000)][constant]\n\n* [GDP, PPP (constant 2005 international $)][ppp]\n\n* [GDP (constant LCU)][lcu]\n\n[constant]: http://data.worldbank.org/indicator/NY.GDP.MKTP.KD[current];\nhttp://data.worldbank.org/indicator/NY.GDP.MKTP.CD[ppp]; http://data.worldbank.org/indicator/NY.GDP.MKTP.PP.KD[lcu];\nhttp://data.worldbank.org/indicator/NY.GDP.MKTP.KN[lcu]\n\n## Automation\n\nDatahub updates this dataset every year automatically.\n\n## License\n\nThis Data Package is made available under the Public Domain Dedication and License v1.0 whose full text can be found at: http://www.opendatacommons.org/licenses/pddl/1.0/",
  "keywords": [
    "GDP",
    "World",
    "Gross Domestic Product",
    "Time series"
  ],
  "last_updated": "2018-01-19",
  "licenses": [
    {
      "name": "ODC-PDDL-1.0",
      "path": "http://opendatacommons.org/licenses/pddl/"
    }
  ]
}
```

Figure 4.1 Fragment of GDP properties described by the Schema.org dataset schema¹²

Metadata terms for an application are often assembled into namespaces from different metadata schemas. Metadata application profiles¹³ 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.¹⁴

A collection of dataset schemas,¹⁵ such as all the datasets in a repository, forms a catalogue. 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.¹⁶ 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 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 rigour 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 knowledge base is, primarily, a SKOS graph that links real-world entities. For example, Wikidata¹⁷ is an effort to develop a knowledge base based on structured data from Wikimedia projects, and VIVO¹⁸ is a knowledge graph of scholarship.

But there are also many stand-alone classification schemes. The Extended Knowledge Organization System (XKOS)¹⁹ 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 et al., 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.²⁰

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.

<i>Version</i>	<i>Sampling: Sampling Procedure, Sampling Unit, Sampling Notes</i>
<i>Study Title</i>	
<i>Alternate Title</i>	<i>Oversampled Group</i>
<i>PIs & Affiliation</i>	<i>Time Method</i>
<i>Funding Agencies</i>	<i>Data Source Type</i>
<i>Summary</i>	<i>Mode of Collection</i>
<i>Subject Terms</i>	<i>Weight</i>
<i>Geographic Coverage Areas</i>	<i>Response Rates</i>
<i>Geographic Representation</i>	<i>Scales</i>
<i>Study Time Periods and Time Frames</i>	<i>Analysis Unit</i>
<i>Collection Notes</i>	<i>Unit of Observation</i>
<i>Study Purpose</i>	<i>Smallest Geographic Unit</i>
<i>Study Design</i>	<i>Data Format</i>
<i>Description of Variables</i>	<i>Restrictions</i>
	<i>Version History</i>

Figure 4.2 ICPSR DDI metadata elements

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

The ICPSR²¹ 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 DDI²² (e.g. Vardigan et al., 2008) is especially well crafted.²³ The DDI codebook saves the exact wording of all the questions and the 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. DDI uses XKOS to provide linked data. Figure 4.2 shows the ICPSR DDI metadata schema.

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 knowledge bases. Moreover, they could be linked with higher-level workflows and mechanisms.

Additional Examples of Repositories

Statistical data collection is a core function of government. Such collections often emphasize social data on employment, criminal justice, and public health, for example. 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. European datasets are maintained in the Consortium of European Social Science Data Archives²⁴ and the European Social Survey.²⁵ Australia has a broad data management initiative, the Australian National Data Service.²⁶ Many US federal governmental datasets are collected at data.gov. In addition, there are many other social survey repositories,²⁷ and many US states and cities have online statistics sites at varying levels of sophistication.

There are also many non-governmental and intergovernmental 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.

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. For instance, massive amounts of geophysical data and related text documents are collected in the EarthCube²⁸ portal. The science.gov portal is maintained by the US Office of Science Technology and Policy. NASA supports approximately 25 different data portals. Each satellite in the Earth Observation System may provide hundreds of streams of data,²⁹ with much common metadata. Likewise, there are massive genomics and proteomics datasets which are accessible via portals such as UniProt³⁰ and the Protein Data Bank,³¹ 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,³² 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 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 JavaScript Object Notation for Linked Data microdata 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 Repository³³ 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 DOIs³⁴ 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. CrossRef³⁵ 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 Health Level Seven International³⁶ and the Kyoto Encyclopedia of Genes and Genomes³⁷ 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, digital repositories must increasingly 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 and Wilson, 2010).³⁸

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 (Controlled Lots of Copies Keep Stuff Safe)³⁹ 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 System⁴⁰ is a policy-based archival management system⁴¹ developed for large data stores. It implements a service-oriented architecture to support best practices established by archivists. Further, audits, such as by the Digital Repository Audit Method Based on Risk Assessment,⁴² may be conducted to assess how well repositories implement trustworthy procedures.

Preservation and provenance metadata schemes such as PREMIS⁴³ and PROV-O⁴⁴ 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 microdata, that is, data for the responses that individuals gave to survey questions.⁴⁵ However, analysis of such microdata 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,⁴⁶ RSpace, and eLab notebooks (ELN) integrate annotations, workflows, raw data, data analysis, and annotations into one environment.

Virtual research environments are typically organized by research communities to coordinate datasets with search and analytic tools. For instance, the Virtual Astronomy Observatory uses Jupyter to provide users with a robust research environment. WissKI⁴⁷ is a platform for coordinating digital humanities datasets which are based on Drupal. Decision support systems 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 Colectica⁴⁸ for DDI and the GSS Data Explorer.⁴⁹ The Amundsen data discovery and metadata engine⁵⁰ uses metadata elements to provide a table explorer. Potentially, interactive visualization tools such as TableLens (Rao and 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 Repository⁵¹ records usage of different metadata terms. A registry may interoperate with editing tools for developers (Gonçalves 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 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 Cube⁵² standard is applied in inter-organizational projects such as EarthCube⁵³. SDMX⁵⁴ 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 modelling 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 and Carroll, 2018). Data files (e.g. Excel files) are just one type of artefact from a research programme; other research objects include workflows. Workflows are a natural fit for describing research methods and analyses (Austin et al., 2017). The Taverna⁵⁵ workflow tool has been used for the MyExperiment⁵⁶ 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 support them. Sequential activities in the management of repositories are often tracked with workflows. For instance, the Generic Statistical Information Model (GSIM)⁵⁷ specifies workflows for the production of datasets by statistical agencies.

Semantic Modelling 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 modelling 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 modelling 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, 2017).

In addition, Allen (2015, 2018) has proposed rich semantic modelling 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 catalogues are implemented with relational databases. Thus, they use SQL and are often characterized by UML class diagrams. Information models (e.g., National Information Exchange Model⁵⁸) 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)⁵⁹. 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 and Gannon, 2017), allowing flexible allocation of computing, networking and storage resources, which facilitates software as a service. 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 analyse the data, which can be a complete digital preservation package.

Highly networked data centres facilitate the Internet of Things 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 for smart cities and have implications for the kinds of questions we may ask about social behaviour.

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.

Acknowledgements

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

References

- Allen, R. B. (2015) Repositories with direct representation. Preprint, arXiv:1512.09070.
- Allen, R. B. (2018) Issues for using semantic modeling to represent mechanisms. Preprint, arXiv:1812.11431.
- Allen, R. B. and Kim, Y. H. (2017/2018) Semantic modeling with foundries. Preprint, arXiv:1801.00725.
- Arp, R., Smith, B. and Spear, A.D. (2015) *Building Ontologies with Basic Formal Ontology*. Cambridge, MA: MIT Press.
- 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. and Carroll, B. (2018) Managing digital research objects in an expanding science ecosystem: 2017 conference summary. *Data Science Journal*, 17. <http://doi.org/10.5334/dsj-2018-016>
- Commission on Evidence-Based Policymaking (2017) *The Promise of Evidence-Based Policymaking*. <https://www.cep.gov/cep-final-report.html>
- Foster, I. and Gannon, D. B. (2017) *Cloud Computing for Science and Engineering*. Cambridge, MA: MIT Press.
- Gonçalves, R. S., O'Connor, M. J., Martínez-Romero, M., Egyedi, A. L., Willrett, D., Graybeal, J. and Musen, M. A. (2019) The CEDAR workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments. Preprint, 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. In M. J. Bates and M. N. Maack (eds), *Encyclopedia of Library and Information Sciences* (3rd edition). Boca Raton, FL: CRC Press.
- Park, J.-R. (2009) Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47, 213–228. doi: 10.1080/01639370902737240
- Rao, R. and Card, S. K. (1994) The Table Lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM, pp. 318–322. doi: 10.1145/191666.191776
- Riley, J. (2017) *Understanding Metadata: What Is Metadata, and What Is It For?: A Primer*. Bethesda, MD: NISO Press.

- 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. doi: 10.2312/re3.008
- Vardigan, M., Heus, P. and 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. and Wilson, A. (2010) *How to Appraise and Select Research Data for Curation*. Edinburgh: Digital Curation Centre.
- 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. (2017) Social mechanisms. In S. Glennan and P. Illari (eds), *The Routledge Handbook of Mechanisms and Mechanical Philosophy*. London: Routledge.

Notes

- 1 The FAIR guidelines have been extended from scholarly texts to datasets (Wilkinson et al., 2016).
- 2 <https://www.ncbi.nlm.nih.gov/pmc/>
- 3 <https://www.aclweb.org/anthology/>
- 4 <https://github.com/allenai/science-parse>
- 5 <https://manpages.debian.org/testing/poppler-utils>
- 6 <https://github.com/explosion/spaCy>
- 7 <https://wiki.dbpedia.org/services-resources/ontology>
- 8 https://en.wikipedia.org/wiki/Category:Statistical_methods
- 9 <https://spacy.io/api/tokenizer>
- 10 <https://rasa.com/docs/nlu>
- 11 <https://nlp.stanford.edu/projects/glove>
- 12 <https://fasttext.cc/docs/en/crawl-vectors.html>
- 13 <https://www.w3.org/TR/shacl/>
- 14 <https://joinup.ec.europa.eu/release/dcat-ap/11>
- 15 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.
- 16 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.
- 17 <https://wikidata.org/>
- 18 <https://duraspace.org/vivo/about/>
- 19 <https://ddialliance.org/Specification/RDF/XKOS>
- 20 <http://www.obofoundry.org/>
- 21 <https://www.icpsr.umich.edu/>
- 22 <http://ddialliance.org>
- 23 DDI is also used for datasets from other organizations such as the National Opinion Research Center (NORC).
- 24 <https://www.essda.eu/>
- 25 <https://www.europeansocialsurvey.org/data/>
- 26 <https://www.ands.org.au/>
- 27 There are additional collections at <http://data.census.gov>, <http://gss.norc.org>, <http://electionstudies.org>, <http://psidonline.isr.umich.edu>, and <http://www.nlsinfo.org>
- 28 <https://www.earthcube.org/>

- 29 <https://pds.nasa.gov/>
- 30 <https://www.uniprot.org/>
- 31 <http://www.rcsb.org/>
- 32 re3data.org
- 33 <https://datadryad.org/>
- 34 <https://datacite.org/>
- 35 <https://www.crossref.org/>
- 36 <https://www.hl7.org/>
- 37 <https://www.genome.jp/kegg/>
- 38 See also, for example, <http://www.dcc.ac.uk/digital-curation/planning-preservation>
- 39 <https://clockss.org/>
- 40 <http://irods.org>
- 41 The policies are based on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) standard; see <https://interparestrust.org/>
- 42 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>
- 43 <http://www.loc.gov/standards/premis/ontology/>
- 44 <https://www.w3.org/TR/prov-o/>
- 45 The term microdata 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.
- 46 <https://jupyter.org/>
- 47 <http://wiss-ki.eu>
- 48 <https://www.colectica.com/>
- 49 <https://gssdataexplorer.norc.org/>
- 50 <https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine-62d27254fbb9>
- 51 <https://mmisw.org/>
- 52 <https://www.w3.org/TR/vocab-data-cube/>
- 53 <https://www.earthcube.org/info/about>
- 54 <http://sdmx.org/>
- 55 <https://taverna.incubator.apache.org/>
- 56 <https://www.myexperiment.org/about>
- 57 <https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model>. GSIM is coordinated with the Common Statistical Production Architecture; see <https://unstats.un.org/unsd/nationalaccount/workshops/2015/gabon/BD/CSPA-ENG.pdf>
- 58 <https://www.niem.gov/about-niem>
- 59 <https://www.openarchives.org/pmh/>

5

COMPETITION DESIGN

Andrew Gordon, Ekaterina Levitskaya,
Jonathan Morgan, Paco Nathan,
Sophie Rand, Sebastian Kohlmeier
and Alex D. Wade

Introduction.....	54
Competition Design.....	56
Data	57
Submission Process	65
Evaluation.....	69
Discussion.....	76
Conclusion.....	78
Appendix: Standardized Metadata, Full Text and Training/Evaluation for Extraction Models (Alex Wade and Sebastian Kohlmeier).....	78
References.....	81
Notes	82

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 programme had funded research teams to develop automated methods that would identify technical emergence using information found in published scientific, technical, and patent literature (Murdick, 2011, 2012), and resulted in recommendation systems such as meta.com (Perrie et al., 2017). Google Dataset Search had developed search technologies that would find datasets in data repositories across the Web (Brickley et al., 2019). Academic social network sites such as ResearchGate and Academia.edu had developed platforms whereby researchers provided feedback about their scientific activities (Ovadia, 2014). And there is a well-established tradition of competitions in computer science, particularly natural language processing (Soborov et al., 2012).

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 programmes') or commitments to reward successful results ('pull incentives') (Kremer and Williams, 2010), 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 (Wright, 1983).

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 (Williams, 2012; Kalil, 2006).

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 (Boudreau and Lakhani, 2009). As open-source software communities show, however, they must be carefully incentivized and nurtured to grow participation (Mateos-Garcia and Steinmueller, 2008) and managed well to maintain resources and quality of output over time (Sadowski et al., 2008).

Specific Challenges

There were a number of challenges associated with developing a natural language processing (NLP) competition that used a corpus consisting of academic publications. One major challenge is that access to scientific publications is often limited and behind a paywall. In addition, there are no existing corpora of empirical publications that have been annotated with the datasets used in the analysis.

Beyond these corpus specific challenges, we found that there are few standards for annotations and existing solutions are not easily reusable. There is a limited community upon which to draw – indeed, a similar project that was 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 (Chapman et al., 2011).

There are some areas in which we could draw heavily on previous expertise, 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 (Soboroff et al., 2012). We also identified 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 (Tong et al., 2018).

We had a broader vision in mind than many competitions as well. We wanted to encourage researchers to go beyond developing a narrow model to solve a problem for a specific corpus. Our goal was to build a community that could develop a generalized model to identify datasets that was not overly dependent on the use of formal

titles of datasets, 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 datasets.

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 labelled corpus of 5000 publications with an additional development fold of 100 publications. Each publication was labelled 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 dataset in our list is mentioned, tie each mention to the appropriate dataset. A separate corpus of 5000 labelled publications was held back to serve as an evaluation corpus. Each team was allowed up to two 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 unlabelled publications and asked to discover which of the datasets from the first phase's data catalogue were used in each publication, as well as to 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 is a corpus that is manually tagged and reviewed for quality, usually for a particular domain and task. Creating one is time-consuming and expensive (Wissler et al., 2014) because it involves selecting a corpus to annotate, then implementing a manual annotation and review scheme.²

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 dataset citation data from the Inter-university Consortium for Political and Social Research (ICPSR) data catalogue,³ then used methods that originated in quantitative content analysis of communication artefacts (Riff et al., 2019) combined with software designed to reduce and simplify the work of human coders to increase reliability (Lewis et al., 2013).

In each of the two phases, competing teams were given text and metadata for 5000 publications and a single set of metadata on 10,348 datasets of interest, shared between the two phases, for use in training and testing their models. A separate set of 5000 publication samples was provided for each phase. The corpus of 10,348 datasets included data maintained by Deutsche Bundesbank and the set of public datasets hosted by the ICPSR. In addition, a single development fold of 100 publications 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.⁴

In each phase, an additional separate set of 5000 publications was held back and used to evaluate the models. After the first phase, the first-phase holdout was also provided to second-phase competitors to serve as additional training and testing data.

In phase 1, both the train-test publications and the holdout publications were broken into 2500 publications each that used one or more of the datasets of interest for analysis, as compiled by ICPSR and Bundesbank staff, and 2500 publications that had not been annotated and had been filtered not to contain data. The dataset citations were captured in a separate dataset 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 unannotated set of 5000 publications to assess their model’s behaviour on uncurated data. The phase 2 holdout of 5000 publications was also unannotated, and evaluation of dataset detection was based on

hand-coded dataset 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 datasets 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 labelled training publications
- 2500 unlabelled/no-dataset training publications
- 100-publication development fold
- 2500 labelled holdout publications
- 2500 unlabelled/no-dataset holdout publications

In phase 1, 5000 publications were provided to participants as a train-test dataset, 5000 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 2500 publications that cited at least one dataset, and 2500 publications that had not been cited by ICPSR as using their data, and had been filtered not to have obvious markers of using data.

The annotated portions of these two sets of publications were drawn from a set of publications provided by the Bundesbank that referenced their data and the publications captured in the ICPSR catalogue annotated as having used a particular dataset 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 5100 total publications with annotated data citations provided to phase 1 participants, the 2550 publications in the train-test corpus (2500) and development fold (50) were randomly selected from the open access set, so they could be distributed freely

to all participants. The 2500 in the holdout were randomly selected from the remainder of the open access set plus those available from Sage. The unannotated publications used in phase 1 were all published by Sage – the 2550 unannotated publications in the train-test corpus (2500) and development fold (50) were open access publications from Sage journals. The 2500 unannotated publications used in the holdout evaluation corpus were sampled from across Sage journal holdings, including non-open access journals.

Publication Dataset - Phase 2

- The main publication corpus for phase 2 of the competition was 10,000 unlabelled publications evenly distributed between six key topic areas (education, health care, agriculture, finance, criminal justice, and welfare), nicknamed the ‘wild corpus’.
- 5000 of these 10,000 were given to teams to work with in phase 2 (randomly selected from within each of the six key topic areas to maintain even distribution across topic areas).
- The other 5000 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 5000 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 3115 non-research articles) that we decided to 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 aim of keeping the distribution across the six topic areas equal (so 1666 randomly selected in two topic areas, 1667 randomly selected in the other four). Then we split the phase 2 corpus to give half to participants and keep half back for evaluation, maintaining an equal distribution between the topic areas within each set of 5000 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 was as follows:

- 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 versus 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 artefacts 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.
- Tables and charts, for example, 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 Datasets

Competitors were provided with two sets of data related to detecting datasets: a catalogue of all of the datasets 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 a subset of these datasets that were actually specifically annotated as having been used for analysis in a given publication.

The dataset catalogue, 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 datasets available from Deutsche Bundesbank. It includes all datasets cited 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.⁵
- `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, datasets, 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 datasets cited in the phase 1 train-test corpus.

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

ICPSR captured when a given dataset 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 dataset–publication pair included in our train-test corpus (see 'Dataset Mention Annotation Process' below). Since we manually created this data, given limited time and resources, we initially only did this work for datasets 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 datasets 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 datasets 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 specifies a single relationship between a dataset 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, the score 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 the 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 dataset in question was not mentioned.

The list of datasets cited in a particular publication is also not exhaustive. There is the possibility that other datasets from our catalogue of datasets 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 datasets. And named datasets not within our catalogue of datasets of interest could also have been used in analysis within a given publication.

Dataset Mention Annotation Process

The ICPSR data contains many explicit ties between publications and datasets 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 dataset citations in our train-test corpus and capture mention text for each. This required human workers to examine each dataset citation in the context of its publication (in other words, to identify how many specific citations there were in all 2500 training publications) – namely, to identify and mark locations in the text where each dataset was referenced.

Because of the manual effort required, we only did this for the 2500 train-test publications that referenced data provided to the teams. We did not manually annotate mention text in the 2500 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–dataset pair (so a given publication would be examined as many times as it had different datasets cited within it).

The ICPSR dataset repository is very fine-grained in definition of a dataset, so each year of an ongoing survey, for example, might have its own dataset. To save time, we eventually created the concept of a dataset family for these types of datasets 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.

Code DataSetCitation mentions

DataSetCitation ID:

Load DataSetCitation

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.

Record Data Set Mentions

Mention ==>

Process Mention | Clear Mention Coding

Mention List

Find in Article Text

Search for:

<== Find in Text

<== Find Words in Text

<== Clear Matches

<== Mention

Data Set Info

Data Set 1 - ANES 1952 Time Series Study

- Title: ANES 1952 Time Series Study

Figure 5.1 The interface of a given publication and a mention-capturing process in the coding tool. The left pane contains the 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 dataset (title, date of collection, formal identifiers), as well as a list of synonyms gathered so far from publications where the dataset is cited

The general process was as follows:

- 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 the status of each, so they could track which they had already completed, and a link for each to the coding page.
- Once the user loaded a particular citation for coding, they were presented with the coding page shown in Figure 5.1, and were asked to follow the coding instructions in the codebook/documentation for the annotation tool.⁶

Coders were instructed to find terms that relate to mentions of the dataset and avoid general synonyms of those terms (e.g., tagging “ANS survey” instead of only “survey”). 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”, “ANS survey” is captured first, and then “ANS survey collected/conducted by X”. 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), rather than trying to capture full text in which the dataset is discussed.

For more details, including an FAQ that provides guidance on specific issues that arose during coding (such as how to deal with datasets that span multiple years), see the content analysis protocol.⁷

In total, a team of five 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 programme). 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. The following considerations motivated 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 tools such as lighttag.io, tag.works, NVivo, Atlas.ti and MAXQDA. Unfortunately, given a tight timeline and relatively complex requirements, we did not 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 of the above tools that did not support distributed workflows.

- For future coding work, we would love to be able to outsource the development of coding tools, and so we are looking at distributed coding applications such as [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 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 uses to classify its publications,⁸ 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 forgo 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 subcategories that we wish we had broken this into: statistical methods, analysis methods, data collection and creation methods, and so on.

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 will 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:

- To 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 development fold to validate processing of standard input formats and creation of required output formats.
- To minimize the installation and configuration work needed on the part of competition organizers to replicate computing environments as part of model submission process.
- To maximize our ability to see and be able to test how each submission environment is set up, and so avoid accepting a black box 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 repository 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 development fold, and shell scripts that implemented the basic steps of packaging, building, running, and testing a model. The Git repository⁹ 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.¹⁰

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', which 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 Docker file with examples of installing OS packages and Python packages 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 repository, for use by the scripts.
- A copy of the development fold, in the standard data folder structure.

The set of scaffold files provided out of the Box folder 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 example files provided 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 two 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 interact 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 ‘Publications’ subsection of the previous section, which 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 four 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 datasets 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 dataset mention detected within each publication, regardless of whether a given dataset is one of the datasets provided in the contest dataset 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 the server was lost.
- The model was downloaded to the server and extracted.
- The submission container was built on the server using the provided Docker file and ‘rcc.sh’, and then the container was run over the development fold to test basic functionality of the container and the model, and to give an estimate of time needed to complete.
- Once the development 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 computer 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 set-up.

To set up the infrastructure in each folder, we created a Git repository that contained all of the files, shell scripts, and templates needed to: configure a new instance of a team folder, for use by competition staff setting up team folders; develop, package and test deployment of a model (participants); and 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 such as 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 were not comfortable that 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,¹¹ and there were also licensing complications that we were not comfortable sorting out. We also needed control over manual evaluation and were concerned that their submission and evaluation system would not support the bespoke nature of our submissions.
- For both, we also simply were not comfortable that we would 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 computer power at a model if it was running too slowly, to get it to complete so we could give the best possible feedback. We not only wanted groups to be able to do preliminary submissions, but also wanted to make sure we could give as much feedback as possible. This led us towards 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 datasets.

Phase 1 Evaluation

Mentions, Methods and Fields

In phase 1, expert social science judges evaluated mentions, methods and fields in two ways: we randomly selected 10 publications to manually examine each team's output against, and made notes of positives and negatives for each team, then ranked the teams within each publication; and 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.

Dataset Citations

To evaluate dataset 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 F_1 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 dataset-publication pairs found in either ICPSR's baseline or the team's output. We created found-or-not (1 or 0) vectors for every dataset-publication pair for the baseline, and for the team. Then, for each dataset-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 the ICPSR on the presence of a dataset, that was counted as a true positive. If the team found a dataset that the ICPSR did not, that was counted as a false positive. If a team missed a dataset that the ICPSR indicated was present, it was counted as a false negative. 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' versus '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/ F_1 scores. For example, even models that figured out that a longitudinal dataset was present sometimes got many false positives and false negatives because they got the years wrong, and models that correctly found ICPSR datasets used in discussion had those counted as false positives because ICPSR had only captured datasets 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 based on what we had learned in phase 1.

Mentions, Methods and Fields

For mentions, methods, and fields in phase 2, we kept the basic strategy of comparing the values created by each team's model in the context of a set of selected publications, and 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 dataset citations), we selected 20 publications for each of our six topic areas of interest (education, health care, agriculture, finance, criminal justice, and welfare) with a few extras (two extra in finance and one 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 datasets (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' (Figure 5.2). Once judges scored all output, we then created rankings based on the sum of each team's scores.

For manual evaluation of dataset 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 dataset 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 one team, two teams, and three teams agreeing data was present proportional to the distribution of each level of agreement in the broader sample:

- 17 from one ($1439/5000 = 0.2878$; $0.2878 \times 60 = 17.268$)
- 20 from two ($1741/5000 = 0.3482$; $0.3482 \times 60 = 20.892$)
- 13 from three ($1080/5000 = 0.216$; $0.216 \times 60 = 12.96$)



Figure 5.2 The interface given to judges to evaluate dataset mentions, research fields, and research methods (The Firefox browser is a trademark of the Mozilla Foundation in the US and other countries)

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

Dataset Citations

Our analysis of dataset citations in phase 2 required a more substantial rethink since we did not have any starting point for the presence or absence of data like the ICPSR corpus. We implemented a method of creating a confusion matrix that could be used to generate precision, recall and F_1 scores more closely aligned with the task we had assigned the teams to implement – finding mentions of data and datasets within publications.

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

- We captured all ‘data references’ within each of those publications using a new human coding protocol. This included external titled datasets 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,¹² 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, to:
 - Collect all mentions that refer to the reference.
 - Decide if the dataset 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’.
 - Capture any broader contextual text in ‘context’, so it could be used to better understand the nature of the ‘data reference’.

- If a dataset title is present, capture it.
- Try looking up the dataset in the database, and if it is there, store its dataset ID.

We tried to capture detailed context on each reference for a couple reasons: first, to make it easier for reviewers of this data to evaluate the quality of each data reference; and second, 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 looked at each data reference related to the selected publications for each team to see if dataset 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. 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 protocol¹³ 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 was not a perfect, formal reference, but also to make sure not to mark things that are obviously unrelated. To help to flesh out this distinction, we gave examples and analogies and training, and we had coders work through a few data references on their own and 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 dataset 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 (Figure 5.3).

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 it to fit his understanding of ‘related to’ or flag it as one he was unsure of and note his thoughts.



Figure 5.3 The interface given to judges to evaluate whether a given team's dataset mentions and citations were related to a given data reference (The Firefox browser is a trademark of the Mozilla Foundation in the US and other countries)

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 in 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 reference, 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 one 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 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. If a team did not have any mentions or citations related to a given data reference, it was counted as a false negative. Then, for the publication, if the team had one or more mentions and/or citations that were not ‘related to’ any data reference, this was counted as a false positive.

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 the Bundesbank has revealed a need to more precisely design how the models work within their Docker container and the APIs they provide so that packaged models implement a more reusable API. For example, to be readily usable 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 reuse, 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 retrained, 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 datasets of interest that are detected specifically without needing to retrain on a full corpus referring to those datasets, but we did not 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 led to collaborations between pairs of sponsors and participants and collective work on making a GSC that could be used to develop better models in the future (a great step towards 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 datasets, the only text available for characterizing a dataset 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 datasets, nor did it include datasets 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 dataset 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 datasets 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 datasets.

In phase 2, our design and evaluation data creation attempted to account for the limitations of phase 1 – to move from just looking at those titled ICPSR and Bundesbank datasets used for analysis and to examine the ways the datasets were discussed in academic papers. 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 eight 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 GSC that could be used to develop better models in the future – an important step towards higher-quality models. The models also ended up being reusable 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 workshop¹⁴ took the next step towards doing so. The third is to further develop the machine learning and NLP 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 (Alex Wade and Sebastian Kohlmeier)

Key challenges when working on an NLP task such as 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 its own approach for disambiguation (e.g. recognizing that A. Smith and Anna Smith are the same author)

or de-duplication of content (clustering preprints 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 corpus¹⁵ developed by the Allen Institute for Artificial Intelligence (AI2), or Microsoft's Academic Graph¹⁶ that makes 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 Unpaywall¹⁷ 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 (3000–5000 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 such as AI2's ScienceParse,¹⁸ which handles key tasks like metadata, section heading and references extraction.

Once the metadata and full text dataset have been created they can be easily versioned and used again in future competitions. For example, if updated metadata is needed it is easy to go back to the original metadata source – for example, by using Semantic Scholar's API¹⁹ to get the latest metadata.

Annotation protocols to produce training and evaluation data

A common approach to machine learning known as supervised learning uses labelled, or annotated, data to train a model in 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 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 among 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 need 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 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 from within 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 overall 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 retrained 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 it 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 a 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 0.80. 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 0.90. 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 F_1 score of 0.847.

If we then make modifications to our model, we can rerun it against the evaluation dataset and see how our F_1 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 F_1 score for each of them. The one with the highest F_1 score has performed the best. Consequently, the quality of the annotations is 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 F_1 score) can be affected by the quality of the data it is evaluated against.

References

- Boudreau, K. J. and Lakhani, K. R. (2009) How to manage outside innovation. *MIT Sloan Management Review*, **50**, 69.
- Brickley, D., Burgess, M. and Noy, N. (2019) in *The World Wide Web Conference*. New York: ACM, pp. 1365–1375.
- Chapman, W. W., Nadkarni, P. M., Hirschman, L., D’Avolio, L. W., Savova, G. K. and Uzuner, O. (2011) Overcoming barriers to NLP for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, **18**(5), 540–543.
- Kalil, T. (2006) *Prizes for Technological Innovation*. Washington, DC: Brookings Institution.
- Kremer, M. and Williams, H. (2010) Incentivizing innovation: Adding to the tool kit. *Innovation Policy and the Economy*, **10**, 1–17 (2010).
- Lewis, S. C., Zamith, R. and Hermida, A. (2013) Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, **57**, 34–52.
- Mateos-Garcia, J. and Steinmueller, W. E. (2008) The institutions of open source software: Examining the Debian community. *Information Economics and Policy*, **20**, 333–344.
- Murdick, D. A. (2011) Foresight and understanding from scientific exposition (FUSE).
- Murdick, D. (2012) *Finding Patterns of Emergence in Science and Technology*. Washington DC: Office of the Director of National Intelligence, Intelligence Advanced Research Projects Activity
- Ovadia, S. (2014) ResearchGate and Academia.edu: Academic social networks. *Behavioral & Social Sciences Librarian*, **33**, 165–169.
- Perrie, J., Hao, Y., Hayat, Z., Colak, R., Lyons, K., Vembu, S. and Molyneux, S. *et al.* (2017) Implementing recommendation algorithms in a large-scale biomedical science knowledge base. Preprint, arXiv:1710.08579.
- Riff, D., Lacy, S., Fico, F. and Watson, B. (2019) *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. New York: Routledge.
- Sadowski, B. M., Sadowski-Rasters, G. and Duysters, G. (2008) Transition of governance in a mature open software source community: Evidence from the Debian case. *Information Economics and Policy*, **20**, 323–332.
- Soboroff, I., Ounis, I., Macdonald, C. and Lin, J. J. (2012) Overview of the TREC-2012 microblog track. In *TREC*, vol. 2012, p. 20.
- Tong, A., Diduch, L., Fiscus, J., Haghpahan, Y., Huang, S., Joy, D., Peterson, K. and Soboroff, I. (2018) Overview of the NIST 2016 LoReHLT evaluation. *Machine Translation*, **32**, 11–30.
- Williams, H. L. (2012) Innovation inducement prizes: Connecting research to policy. *Journal of Policy Analysis and Management*, **31**, 752–776.

- Wissler, L., Almashraee, M., Díaz, D. M. and Paschke, A. (2014) in *IEEE GSC*.
- Wright, B. D. (1983) The economics of invention incentives: Patents, prizes, and research contracts. *American Economic Review*, 73, 691–707.

Notes

- 1 The public description of the competition is available at <https://coleridgeinitiative.org/richcontextcompetition>
- 2 Wissler et al. (2014) outline options for decreasing the cost, including starting with a ‘silver standard corpus’ created using chained machine learning models and annotation via crowdsourcing.
- 3 <https://www.icpsr.umich.edu/icpsrweb/>
- 4 For details on the metadata provided for each type of data, see <https://github.com/Coleridge-Initiative/rich-context-competition/wiki/Dataset-Description>
- 5 <https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/subject>
- 6 https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit
- 7 https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit
- 8 <https://sagepub.libguides.com/research-methods>
- 9 <https://github.com/Coleridge-Initiative/rich-context-competition>
- 10 <https://github.com/Coleridge-Initiative/rich-context-competition/wiki>
- 11 Our evaluation steps were fuzzy, so could not 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 do not have the data yet.
- 12 <https://docs.google.com/document/d/1aFPEtT4hd93kcsOEzocyB6-a4Hu8WcemKTld-98Q25k/edit/#heading=h.f3u3kdbg87s4>
- 13 https://docs.google.com/document/d/1Hi13N6gfRz9nfwCoUQrey8v_ozY7fKHMtH-V4GgX2ys/edit
- 14 <https://coleridgeinitiative.org/richcontextworkshop>
- 15 <http://api.semanticscholar.org/corpus/>
- 16 <https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema>
- 17 <https://unpaywall.org/data-format>
- 18 <https://github.com/allenai/spv2>
- 19 <http://api.semanticscholar.org>

6

FINDING DATASETS IN PUBLICATIONS: THE ALLEN INSTITUTE FOR ARTIFICIAL INTELLIGENCE APPROACH¹

Daniel King, Waleed Ammar, Iz Beltagy,
Christine Betts, Suchin Gururangan and
Madeleine van Zuylen

Introduction.....	84
Methods	84
Results.....	88
Conclusion.....	90
Acknowledgements.....	91
References.....	91
Notes	92

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,² 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 artefacts 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. Semantic Scholar covers more than 180 million papers from all domains of science.

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: extracting the datasets used in publications; predicting the field of research of publications; and 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: citation prediction (extraction and linking to a provided knowledge base of *known datasets*) and mention prediction (extraction of both *known and unknown* dataset mentions).

Provided Data

The provided knowledge base of known datasets includes approximately 10,000 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 5000 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 labelled mentions in the knowledge base.

We provide a high-level overview of our approach in Figure 6.1. First, we used an NER model to predict dataset mentions. For each mention, we generated 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 used a binary classifier to predict which of these candidates is a correct dataset extraction.

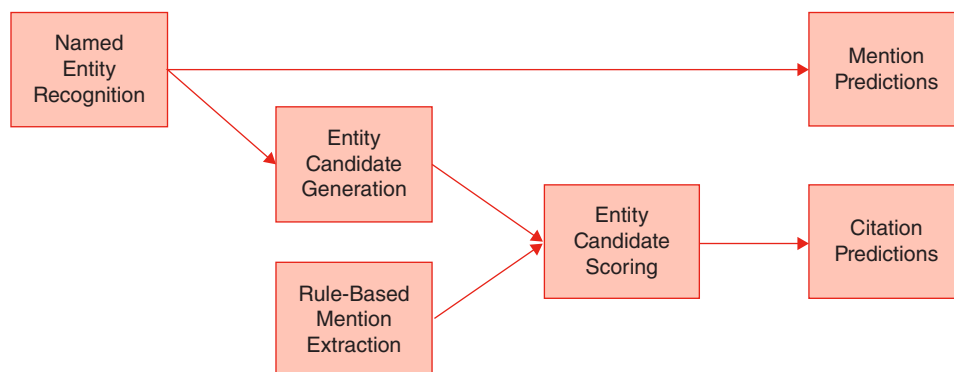


Figure 6.1 A high-level overview of our approach to dataset mention detection and linking

Next, we describe each of the subcomponents 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, on both the provided development fold and our own development fold that we created. High recall and low precision were 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 bidirectional long short-term memory (bi-LSTM) model with a conditional random field (CRF) decoding layer, similar to Peters et al. (2018), and implemented using the AllenNLP framework (Gardner et al., 2018).³ 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 resulted in noisy labelled 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), and 16-dimensional

character embeddings with 64 convolutional neural network (CNN) filters of sizes (2, 3, 4). The CNN character encoder outputs 128-dimensional vectors. We optimize model parameters using ADAM (Kingma and Ba, 2015) 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 term frequency-inverse document frequency (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 XGBoost⁴ 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 (e.g., 2018) appears both in the mention context and in the dataset title (binary).
- Mention length (integer).
- Mention sentence length (integer).
- 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 (integer).

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 on a machine with eight 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 75,000 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 seven fields, while the more granular one (L1) has 32. Fields associated with fewer than 100 papers were excluded.

Methods

For each level, we trained a bi-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 each 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 kinds 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 NER, and using open information extraction 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 sorted datasets by the number of papers they appeared in, then processed one dataset at a time. For each dataset, we chose one of the train, development or test splits at random and added the dataset to it, along with all papers which mention that dataset. When there was 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 were added to the same split as d_1 . For any further conflicts, we preferred 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, deep contextual embeddings like ELMo could be leveraged.

Table 6.1 NER precision, recall and F_1 performance (%) on the development and test sets

	Precision	Recall	F_1
Development set	53.4	50.3	51.8
Test set	50.7	41.8	45.8

Table 6.2 End-to-end precision, recall and F_1 performance (%) for citation prediction on the development set provided in phase 1 of the competition

	Precision	Recall	F_1
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

	Precision	Recall	F_1
+ development set examples	77.0	47.0	58.3
+ NER mentions	56.3	62.0	59.0

Table 6.3 End-to-end precision, recall, and F_1 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

	Precision	Recall	F_1
Phase 1 holdout	35.7	19.6	25.3
Phase 2 holdout	39.6	18.8	25.5

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 (F_1) of 38.4%. Adding $p(d \mid m)$ and $p(m \mid d)$, which are the probability of entity given mention and probability of mention given entity, improves the performance (by 2.3% to 40.7%). Year matching helps disambiguate between different datasets in the same series, which was found to be a major source of errors in earlier models (by a further 2.8% to 43.5%).

Aggregating mentions for a given dataset, adding mention and sentence length features, adding a feature indicating whether a mention is an acronym, and further hyperparameter tuning improve the results (by 12.0% to 55.5%). Adding examples in the development set while training the model results in further improvements (a further 2.8% to 58.3%). Finally, adding the NER-based mentions significantly improves recall at the cost of lower precision, with a positive net effect on F_1 score (raising it by 0.7% to 59.0%).

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 hyperparameters, evaluated on a held-out development set of papers from the MAG. 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 the MAG for this problem is that our model cannot find new fields of research, and is limited to those provided by the 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 a yield of 1.5 methods per paper on the manually inspected subset of papers. Similarly to research area prediction, the main limitation 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 labelled 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 have a high-confidence evaluation dataset, and to look for other data sources that are similar enough to the task at hand to be useful. 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 reproducibility, 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.

Acknowledgements

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

- Gardner, M., Grus, J., Neumann, Mark., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M. E., Schmitz, M. and Zettlemoyer, L. S. (2018) AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 1-6.
- Kingma, D. P. and Ba, J. L. (2015) Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations 2015 (ICLR)*.
- Pennington, J., Socher, R. and Manning, C. D. (2014) GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Red Hook, NY: Curran.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. S. (2018) Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*:

Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA: Association for Computational Linguistics, pp. 2227–2237.

Shen, Z., Ma, H. and Wang, K. (2018) A Web-scale system for scientific knowledge exploration. In F. Liu and T. Solorio (eds), *ACL 2018: 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics, pp. 87–92.

Notes

- 1 The code for the submission can be found on GitHub (<https://github.com/allenai/coleridge-rich-context-ai2>). There is a readme with additional documentation at this GitHub repository.
- 2 <https://www.semanticscholar.org/>
- 3 https://github.com/allenai/allennlp/blob/v0.9.0/allennlp/models/crf_tagger.py
- 4 <https://xgboost.readthedocs.io/en/latest/>

7

FINDING DATASETS IN PUBLICATIONS: THE KAIST APPROACH

Haritz Puerto San Roman, Giwon Hong,
Minh-Son Cao and Sung-Hyon Myaeng

Non-Technical Overview	94
Literature Review	94
What We Did.....	94
What Worked and What Did Not.....	99
Summary of Results and Caveats	100
Lessons Learned and What We Would Do Differently	103
Conclusion.....	103
What Comes Next	103
Acknowledgements.....	104
Appendix: Description of the Code and Documentation.....	104
References.....	104

Non-Technical Overview

The KAIST approach for retrieving datasets is to generate questions about datasets such as *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, one of which should be the name of the dataset. To remove incorrect candidate answers, we 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 usually created by organizations.

For research field retrieval, we proposed to compare publications with Wikipedia articles to discover the research fields. First, we crawled Wikipedia articles corresponding to the list of research fields. Then we retrieved the research fields of the publications by measuring the similarity between the papers and the crawled Wikipedia documents. For example, we 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*. We proposed to use the TF-IDF as similarity measure, which is based on term frequency and document frequency, but others could be applied too.

For research method retrieval, we modelled the task as a named entity recognition (NER) problem. We 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 from publications. Ghavimi et al. (2016) 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 science 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 We Did

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

Datasets Retrieval

The approach followed to retrieve dataset names is based on machine reading for 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 paragraph–query pair 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 what follows, the MRQA, query generation, and entity typing models are explained in detail.

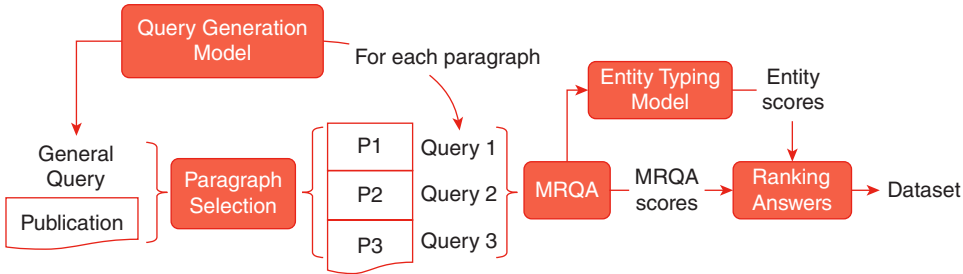


Figure 7.1 Overall architecture for dataset retrieval

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 (Clark and Gardner, 2017). It uses Bidirectional Gated Recurrent Units (Bi-GRU, Cho et al., 2014), bi-attention, and self-attention mechanisms. 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 for training MRQA models (Rajpurkar et al., 2016). For details about the implementation and computing resources to train it, refer to the original publication.

Our hypothesis was that MRQA models do not need the full publication to find datasets. Rather, they only need to process the paragraph where the answer appears. The model used in this work stands out in the MRQA literature 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 such as surveys, datasets or studies. Therefore, we 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, we used a query generation model proposed by (Yuan et al., 2017) 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, we 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 of a dataset. Because of this, these query terms have the discriminatory power to retrieve dataset mentions since queries are generated to extract mentions, and the query terms do not appear in the sentences without dataset mentions.

This list is used twice in our 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 et al., 2018) can predict a set of free-form 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 10,000 different entity types among which *dataset* is included. However, after a few experiments, we 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 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, we used the pre-trained model that was released with the original publication. For details about the implementation and computing resources to train it, 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, we were able to create a neural network classifier to filter out candidate answers.

The classifier has the following architecture:

- 1 Input size: 10,332 (10,331 labels from ultra-fine entity typing and one from the Document QA score)
- 2 One hidden layer with 50 neurons
- 3 Output size: 2

The training set consists of 25,172 examples and the test set of 6293 examples obtained from the training set provided by the competition. To train the network, the Adam optimizer (Kingma and Ba, 2015) was used with cross-entropy as the loss function.

Research Fields Retrieval

Our approach to obtaining 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 in Figure 7.2. The leaf nodes of that hierarchy were searched in Wikipedia to retrieve specific research fields instead of general ones. For example, we were aiming to retrieve *Neurosurgery* instead of *Medicine*. Then, using Scikit-learn (Pedregosa et al., 2011), a TF-IDF matrix of all the publications and Wikipedia articles of research fields was 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.

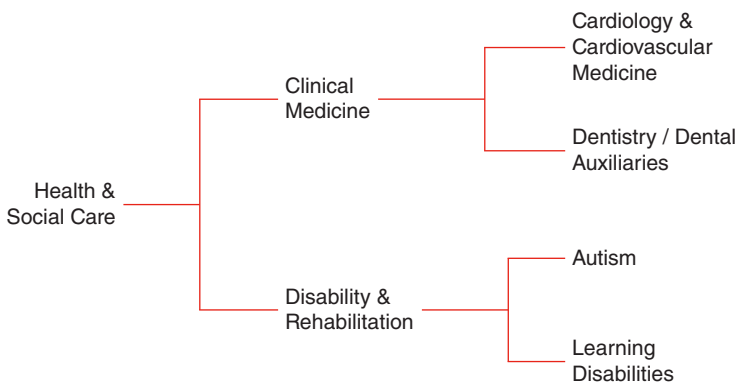


Figure 7.2 Example of research fields hierarchy

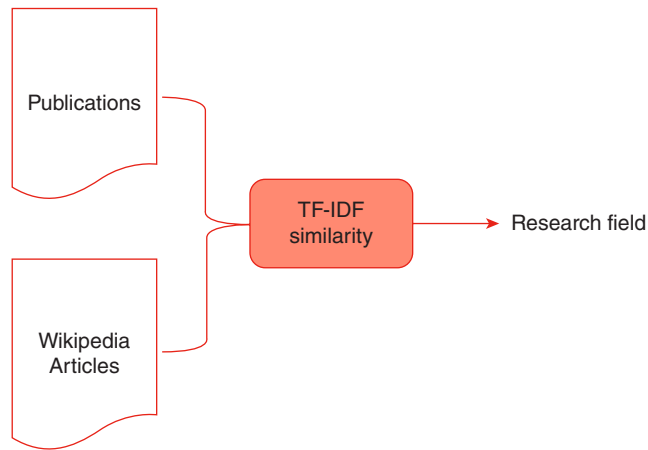


Figure 7.3 Overall architecture for research fields retrieval

Research Methods Retrieval

We modelled the research methods retrieval task as an NER problem. Research methods are considered to be named entities and, because of this, they can be tagged with a research method label (RM) instead of common NER labels such as *location* and *people*. Figure 7.4 shows the main architecture of the model proposed by Lample et al. (2016) and used in this task.

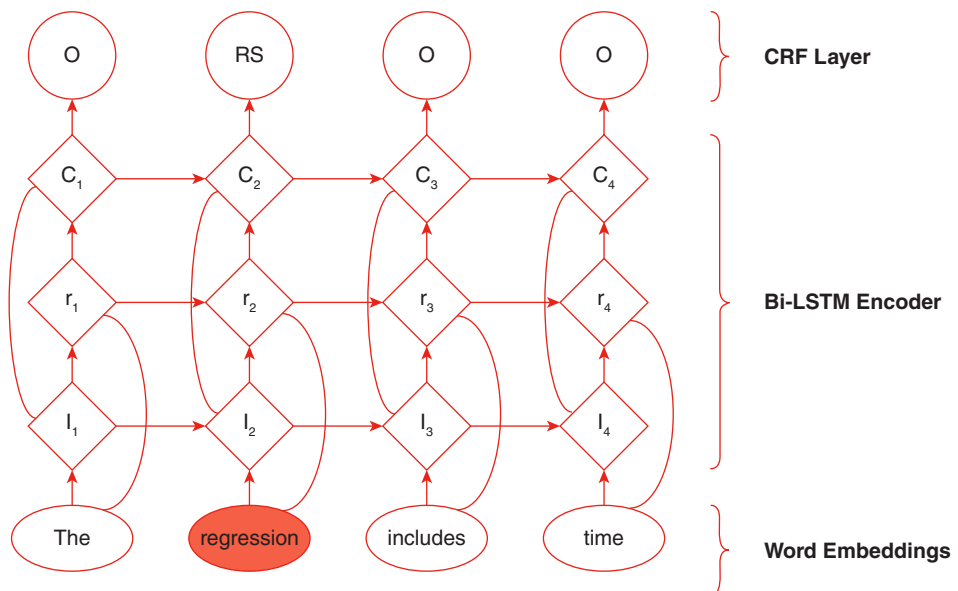


Figure 7.4 Architecture of the named-entity recognition tool for research method retrieval

The representation of a word using the model is obtained considering its context. We assumed that research methods have dependencies and constraints with words that appear in their surrounding context. Therefore, the conditional random field (CRF; Lafferty et al., 2001) 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 (Tjong Kim Sang and De Meulder, 2003), using IOB tagging (Inside, Outside, Beginning) to train the model. Every token was labelled as B-RM if the token was the beginning of a research method, I-RM if it was inside a research method but not the first token, or O otherwise.

Training the model took approximately 1 day using an AMD Ryzen 7 2700 CPU, an Nvidia GeForce GTX 1050 Ti GPU, and 8GB RAM.

What Worked and What Did Not

We tried different ideas to extract dataset names. First, we tried to extract the dataset names using hand-crafted queries in the MRQA model. However, we noticed that these manually generated queries do not have sufficient discriminative power. Therefore, we tried to generate a general query with enough discriminative power to retrieve datasets' names. To this end, we converted the sentences containing the dataset into queries and then clustered the converted queries to get general queries. However, we found that each of the resulting clusters did not reflect the semantics of the desired general queries. Hence, we 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 negatively affected 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, we were able to improve the recall using the query generation module without sacrificing precision.

We 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 our analysis, we stated that the heuristics that we used to extract them generated noise that affected the performance. For example, *total*, *variable* and *funding* were detected as section names but clearly are not.

Finally, our idea of comparing publications with Wikipedia articles to retrieve research fields yielded successful results. But on the other hand, our first idea of retrieving 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 the lack of discriminative power of the most common words that co-occur with the research methods. Therefore, we tried to model it as an

NER problem, considering each research method that appeared in a publication as a named entity. By modelling the problem in this way, we could use existing NER models to extract research methods from publications. However, this approach did not achieve satisfactory results either.

Summary of Results and Caveats

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

Datasets Retrieval

To analyse the effects of the query generation and entity typing modules, we performed analyses on 100 publications from the development 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 analysed 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 the *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 analysed publications.

	1134.txt	153.txt	143.txt
Datasets (Ground Truth)	<ul style="list-style-type: none">• National Comorbidity Survey (NCS)• British Psychiatric Morbidity Survey	<ul style="list-style-type: none">• Micro Database Direct Investment (MiDi)	<ul style="list-style-type: none">• Micro-level German International Trade in Services (ITS) Data• Micro Database Direct Investment (MiDi)
Model Output	<ul style="list-style-type: none">• National Comorbidity Survey	<ul style="list-style-type: none">• None	<ul style="list-style-type: none">• MiDi• The Balance of Payments Statistics

Figure 7.5 Results from Document QA only. Correct answers from the model in red

These results were expected due to the difficulty of defining general queries as explained in section *Query 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 analysed 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 the *153.txt* publication contains several answers including the correct one, *Micro Database Direct Investment*. Therefore, this and the retrieval of more than 2000 answers from the list of analysed datasets proves that the query generation module improves the recall of the entire dataset retrieval model. On the other hand, compared to the Document QA only setting, there is a considerable number of incorrect candidate answers. For instance, in Figure 7.6, *empirical*, *Table 1*, and *Section 4* are not dataset mentions.

We believe that the reason for this noise is that some query terms may cause the retrieval of incorrect 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 incorrect 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.

	1134.txt	153.txt	143.txt
Datasets (Ground Truth)	<ul style="list-style-type: none"> National Comorbidity Survey (NCS) British Psychiatric Morbidity Survey 	<ul style="list-style-type: none"> Micro Database Direct Investment (MiDi) 	<ul style="list-style-type: none"> Micro-level German International Trade in Services (ITS) Data Micro Database Direct Investment (MiDi)
Model Output	<ul style="list-style-type: none"> National Comorbidity Survey British Psychiatric Morbidity Survey NCS Table 1 Psychosis 	<ul style="list-style-type: none"> Financial services FDI data Empirical Deutsche Bundesbank (the German central bank) Micro Database Direct Investment //go.worldbank.or/SNUSW978P0 Mixed logit model 	<ul style="list-style-type: none"> Empirical ITS data Collective reports Transactions below the reporting limit of €12,500 Section 4 4 2.1 Micro Data

Figure 7.6 Results from Document QA + query generation module. Correct answers from the model in red

Document QA + Query Generation Module + Entity Typing Module

Figure 7.7 shows the results of the analysed publications using Document QA, the query generation module and the entity typing module. Although the entity typing module might not remove all the incorrect 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 development set of the first phase of the competition.

	1134.txt	153.txt	143.txt
Datasets (Ground Truth)	<ul style="list-style-type: none">• National Comorbidity Survey (NCS)• British Psychiatric Morbidity Survey	<ul style="list-style-type: none">• Micro Database Direct Investment (MiDi)	<ul style="list-style-type: none">• Micro-level German International Trade in Services (ITS) Data• Micro Database Direct Investment (MiDi)
Model Output	<ul style="list-style-type: none">• National Comorbidity Survey• British Psychiatric Morbidity Survey• NCS• The National Comorbidity Survey	<ul style="list-style-type: none">• Micro Database Direct Investment	<ul style="list-style-type: none">• 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. Correct answers from the model in red

Research Fields Retrieval

We randomly selected 20 publications from the training set of the first phase. We could test our 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*. One weak point of the model is that 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

Twenty random publications were selected and labelled from the training set of the second phase and the results were not satisfactory. The model was able to find proper research methods for 12 publications out of the 20. For example, the model successfully detected 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 incorrect answers such as *reviews describe*, *composite materials*, and *detailed databases*.

Lessons Learned and What We Would Do Differently

After the completion of this project, we realized that some steps could have been done differently and would have led to better results. For example, we focused a lot on the model creation, but we 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 our hypothesis of modelling dataset retrieval as a question answering task was correct. However, at some point during the project, we should have changed it to another model with state-of-the-art performance. Also, we used symbolic queries for the MRQA model. But since they generate 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 such as BM25, a ranking function used by search engines whose performance is better than TF-IDF. Finally, for research methods, they should have further analysed the dataset to use more suitable techniques such as unsupervised NER instead of supervised NER.

Conclusion

We 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. The approach is flexible because it allows the retrieval of new datasets that are not in the training set. In addition, the model does not need to be trained on a dataset discovery task. We achieved good results even though we used a pre-trained model on SQuAD, a dataset for question answering using Wikipedia pages.

Our approach to retrieving research fields is simple, fast to compute, and powerful. It can retrieve specific research fields with high precision. On the other hand, our approach to retrieving research methods did not achieve as good results as in the other task. The main problem was that we did not tackle the noise 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. We hypothesize 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 (Rajpurkar et al., 2016). Therefore, according to our hypothesis, two publications using SQuAD will presumably 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 such as SQuAD and TriviaQA (Joshi et al., 2017). 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 SQuAD and TriviaQA. Moreover, it would be able to recommend research fields with a lack of datasets to researchers interested in creating new datasets.

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

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2013-0-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services) 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>

References

- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 1724-1734.
- Choi, E., Levy, O., Choi, Y. and Zettlemoyer, L. (2018) Ultra-fine entity typing. In F. Liu and T. Solorio (eds), *ACL 2018: 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of System Demonstrations (Volume 1, Long Papers)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 87–96.

- Clark, C. and Gardner, M. (2017) Simple and effective multi-paragraph reading comprehension. Preprint, arXiv:1710.10723.
- Ghavimi, B., Mayr, P., Vahdati, S. and Lange, C. (2016). Identifying and improving dataset references in social sciences full texts. Preprint, arXiv:1603.01774.
- Joshi, M., Choi, E., Weld, D. S. and Zettlemoyer, L. (2017) Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. Preprint, arXiv:1705.03551.
- Kingma, D. P. and Ba, J. L. (2015) Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations 2015 (ICLR)*.
- Lafferty, J., McCallum, A. and Pereira, F. C. N. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, pp. 282–289.
- Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016) Neural architectures for named entity recognition. Preprint, arXiv:1603.01360.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016) SQuAD: 100,000+ questions for machine comprehension of text. Preprint, arXiv:1606.05250.
- Tjong Kim Sang, E. F and De Meulder, F. (2003) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, Volume 4. Morristown, NJ: Association for Computational Linguistics, pp. 142–147.
- Yuan, X., Wang, T., Gulcehre, C., Sordoni, A., Bachman, P., Subramanian, S., Zhang, S. and Trischler, A. (2017) Machine comprehension by text-to-text neural question generation. Preprint, arXiv:1705.02012.

8

KNOWLEDGE EXTRACTION FROM SCHOLARLY PUBLICATIONS: THE GESIS CONTRIBUTION TO THE RICH CONTEXT COMPETITION

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

Introduction.....	108
Approach, Data and Preprocessing	109
Dataset Extraction	111
Research Method Extraction	114
Research Field Classification	120
Discussion and Limitations	122
Conclusion.....	124
Acknowledgements.....	124
References.....	124
Notes	126

Introduction

GESIS – 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 Knowledge Technologies for the Social Sciences (WTS)² department is responsible for developing all digital services and research data infrastructures at GESIS and aims to provide integrated access to social science data and services. Next to traditional social science research data, such as surveys and census data, an emerging focus is to build data infrastructures able to exploit novel forms of social science research data, such as large Web crawls and archives.

Research at WTS³ addresses areas such as information retrieval, information extraction and natural language processing (NLP), semantic technologies and human–computer interaction and aims to ensure access and use of social science research data in accordance with 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 science 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 science resources and to understand their inherent relations.

Previous work at WTS/GESIS addressing such challenges applies NLP and machine learning techniques to, for instance, extract and disambiguate mentions of datasets⁴ (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 *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* platform⁶ (Krämer et al., 2018), which enables search across a vast number of social science 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 competition and are looking forward to continuing 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

the quantitative results of our initial contribution throughout the first phase showed significant room for improvement, the qualitative assessment, conducted by four judges on a sample of 10 documents, underlined the potential of our approach.

This chapter describes our approaches, techniques, and additional data used to address all three competition tasks. As described below, we decided to follow a module-based approach where each module or the entire pipeline can be reused. The rest of the chapter is organised as follows. We begin by providing an overview of our approach, background data and preprocessing steps, and then describe our approaches in more detail, including results towards each of the tasks. Finally, we discuss our results and provide an overview of future work.

Approach, Data and Preprocessing

This section describes the external data sources we used as well as our preprocessing steps.

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 alone or as part of a data-processing pipeline. Figure 8.1 depicts the modules developed and their dependencies. Here, the upper three modules (in grey) describe the preprocessing steps (see Chapter 2). The lower four modules (in red) are used to generate the output in a predefined format as specified by the competition.

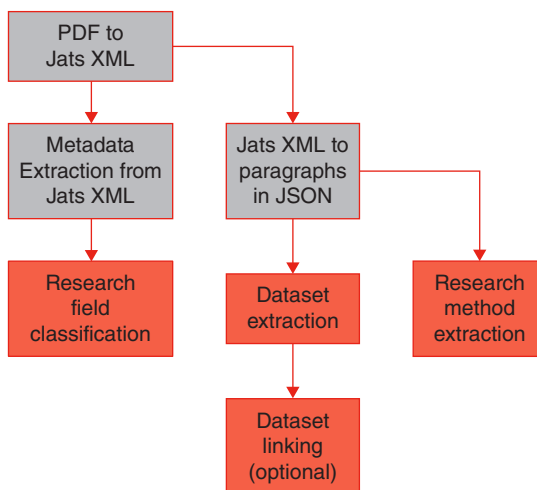


Figure 8.1 An overview of the individual software modules described in this chapter and their dependencies: our preprocessing pipeline (grey); the three main tasks of the RCC (red)

The preprocessing 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, methods and fields (see below). First, a named entity recognition (NER) module is used to find dataset mentions. This module uses a supervised approach trained on a weakly labelled 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 using a named entity recognition and linking module with incorporated word embeddings and lexical resources. To identify 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 Repository⁸ (SSOAR). We tried different classifiers and selected the best-performing one, a classifier based on fastText,⁹ that is, a neural-net-based approach with high performance (Joulin et al., 2017).

After the first phase, each team received feedback from the organizers of the RCC. The feedback was twofold, 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. However, our approach was found convincing regarding the quality of results. The qualitative feedback was based on a random sample of 10 documents given to four judges. The judges were asked to manually extract dataset mentions. Then the overlap between their dataset extractions and the output of our algorithm was calculated. Other factors that judges took into consideration were specificity, uniqueness, and multiple occurrences of dataset mentions. As for the extraction of research methods and fields, no ground truth was provided; these tasks were evaluated against the judges' expert knowledge. Similarly, to the extraction of dataset mentions, specificity and uniqueness were considered for these two tasks. The feedback our team received was overall positive.

External Data Sources

To develop our algorithms, we utilized two external data sources. For the discovery of research methods and fields, we resorted to data from SSOAR.¹⁰ 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 Classification¹²) 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-PMH¹³ 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.

Preprocessing

Although the organizers of the RCC offered plain texts for each publication, we decided to build our own preprocessing 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. *Cermin*¹⁴ (Tkaczyk et al., 2015) transforms the files into XML documents using the Journal Article Tag Suite (Jats).¹⁵ For the competition we identified two interesting elements of the Jats XML format, namely, `<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 preprocessing, 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.

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

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

Dataset Extraction

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 extracting and 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 for citing datasets. Although there are proposed standards for dataset citation in full texts, researchers still ignore or neglect such standards (see, for example, 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. Example 1 illustrates the use of abbreviations that are known mainly in the author's research domain. Example 2 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) and *ISSP* (dataset).*

Example 2: We used *WHO data* (dataset) from 2001 to estimate the spreading degree of AIDS in Uganda.

We treat the problem of detecting dataset mentions in full text as an 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).

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'); and (iii) a vague reference (e.g. 'the monthly statistic'). With all three types, the NER task faces special challenges. In the first case, the dataset name used can vary in different publications. For instance, one publication cites the dataset as 'National Health and Nutrition Examination Survey', while another could use the words 'Health and Nutrition Survey'. In the case where abbreviations are used, a disambiguation problem occurs, for example, in 'WHO data'. WHO may describe the World Health Organization or the White House Office. If 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 dealt with this lack of ground truth data.

Phase 1 Approach

Missing ground truth data was the main problem to be dealt with during this competition. Supervised learning methods for dataset mention extraction from texts are not applicable without the identification of external training data or the creation of useful labelled training data from information given by the competition. Because of the lack of existing training data for the task of dataset mention extraction we resorted to the list of dataset mentions and publication pairs provided and reannotated the particular sentences in the publication text. A list of dataset-identifying words was provided by the competition for some of the known links between publications and datasets. These words represent the evidence of the linkage between publication and datasets and were extracted from the publication text. In the course of reannotation, we searched for each of the identifying words in the corresponding publication texts. For each match, we annotated the occurrence in our raw text and used these annotations as ground truth. As described in the preprocessing section, our units for processing the publication text are paragraphs. The reannotated 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 reannotation was then used to train spaCy's neural-network-based NER model.¹⁶ We created a holdout set of 1000 publications and a training set of size 4000. Afterwards, we trained our model with the paragraph as a sampling unit. In the training set, 0.45% of the paragraphs contained mentions. For each positive training example, we 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.

Table 8.2 Performance of phase 1 approach for dataset extraction

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

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. This means that even if we could not match the dataset mention in a text exactly, we can find the right context with very high precision.

Phase 2 Approach

In the second phase of the competition, 5000 additional publications were provided by RCC. We extended our approach to consider the list with dataset names supplied by the organizers and reannotated the complete corpus of 15,000 publications in the same manner as in phase 1 to obtain training data. This time we split the data into 80% for training and 20% for test.

Evaluation

We resorted to the same evaluation metrics as in phase 1. However, we calculated 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 1, there is also a smaller difference between the precision and recall values for the strict and partial version of the metrics.

Table 8.3 Performance of phase 2 approach for dataset extraction

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

Research Method Extraction

Task Description

Inspired by recent work by 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*, *model*, *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 entity 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 3: *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 examination of the Sage ontology given by the RCC as a sample of how research method terms might look.

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

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

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.

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 the *linking problem*. Note that part of this problem also results from PDF-to-text conversion, which is error-prone. Dealing with incomplete knowledge bases, that is, *handling out-of-vocabulary 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, resulting in insightful co-occurrence statistics, providing additional detail directly related to entity resolution, and helping to assess the *relevance of terms* by means of a score.

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 preprocessing, classification, linking and term filtering.

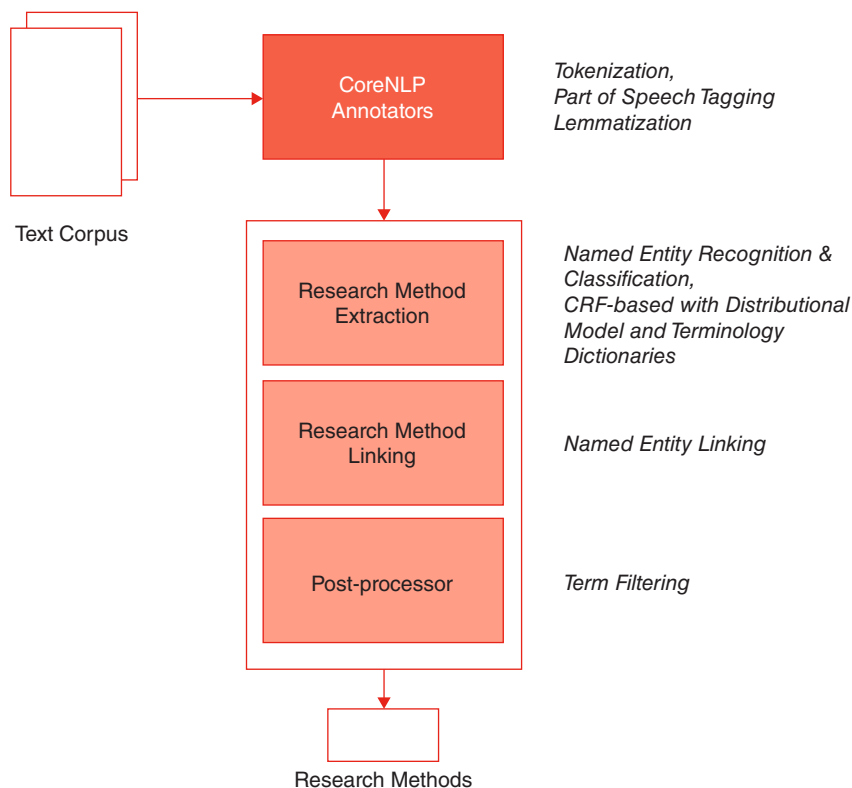


Figure 8.2 Overview of the entity extraction pipeline

We envision the task of finding entities in scientific publications as a sequence labelling 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 conditional random field (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 the *Sage Encyclopedia 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 unit-hood, while scoring is achieved by calculating a relevance score based on TF-IDF (see Table 8.6).

Our research experiments are based on publications from SSOAR²⁰ as well as the train and test data of the RCC corpus.²¹ Our work extends previous work on this topic (see Eckle-Kohler et al., 2013) in various ways. First, we do not limit our study to abstracts, but use the entire full text. 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 a 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 Corpus²² using Gensim²³ with the skip-gram model (see Mikolov et al., 2013) and a pre-clustering algorithm.²⁴

Features

The features incorporated into the linear chain CRF are shown in 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, orthographics, gazetteer, part-of-speech (POS) tags, along with word clustering have been used.

Table 8.5 Features used for NER

Type	Features
Token unigrams	$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, \dots$
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

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 SSOAR. More concretely, we carried out the following steps to extend Sage as an offline 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 term 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

A total of 26,082 candidate terms have been recognized and classified by our pipeline and manually inspected to find synonyms and related words that could be linked to Sage, and to build a post-filter for incorrectly classified terms. Moreover, abbreviations have been extracted using the algorithm of Schwartz and Hearst (2003). In this way, a named entity gazetteer could be built and is used at run-time. It comprises 1111 terms from Sage and 447 terms from the 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 offline 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.

Table 8.6 Most relevant terms from Sage by semantic type

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

Definition of a Relevance Score

The 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), preceded or followed by modifiers such as adjectives (JJ), participles (VB*) or cardinal numbers (CD), complemented by word-shape features.

In order to find out if the candidate term also fulfils 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 are 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 full text papers from SSOAR and 10 documents from the RCC, all randomly selected from a holdout 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: the judges reviewed 10 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

Research Field Classification

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 from the publications. This approach sees the task as a keyword extraction task and considers the terms detected 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 supervised manner. The benefit of the second approach is that the classification scheme to describe the research field can be defined by domain experts. 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 n labels $\{C_i = \emptyset \cap c_1 \dots c_n \mid c_n \in L\}$.

n denotes the number of labels from L describing the research field 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 \emptyset .

Our Approach

Since we did not receive any gold standard for this task during the competition we decided to make use of external resources. We decided to use an external labelled dataset

to train a text classifier which is able to predict one or more research labels for a given abstract of a publication.

Because publications given throughout the competition belong to the domain of social sciences we considered training data from the same domain. Namely from 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 SSOAR annotated data contains four different annotation schemes for research-field-related information. Having reviewed 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 exhaustive description of the data used was given earlier in this chapter.

Preprocessing 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 abstracts 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 counted 22,453 English abstracts with valid classification after filtering. Due to the unequal distribution of labels in the dataset, we needed to guarantee 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. To create train and test sets, 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. The used implementation was written by the authors of this paper.²⁹ The arguments to use this model were the speed compared to a more complex neural net architecture and that the performance was comparable to the state of the art (e.g. Wang et al., 2018). The model was trained with learning rate 1.0 for 150 epochs. Also, the negative sampling parameter was set to 25.

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 multilabel 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 F_1 score is achieved. By increasing the

threshold from this point, the micro-precision score is increasing, but the micro-recall is falling. By decreasing the threshold, these trends are inverted. Also, the default threshold of 0.5 does not look promising. In spite of a micro-precision score of about 0.75, we have a problem with the very high number of items without any prediction. In respect of this observation it is advantageous to select a lower threshold in a productive environment. The ‘without prediction’ curve 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’ 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 is correct. This value decreases with increasing threshold similar to the recall metric. The final micro F_1 value on the test set for our model and a selected threshold of 0.1 is 0.56 (precision 0.55, recall 0.56).

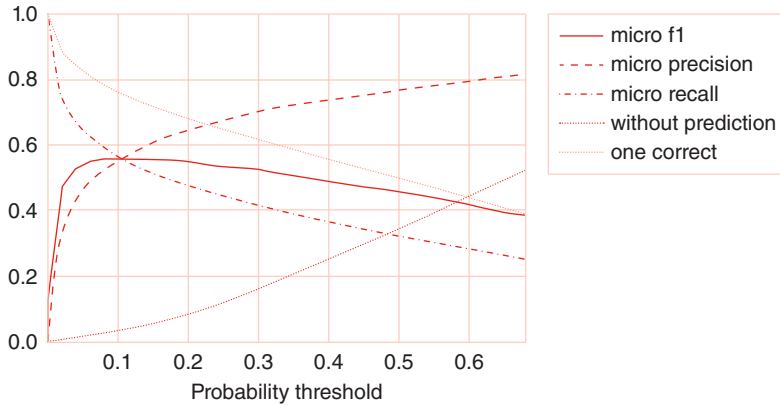


Figure 8.3 Performance for different selected probability thresholds (validation set)

Discussion and Limitations

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 holdout dataset from our automatically created dataset. The recall may be biased, given that our training and test datasets are biased towards known datasets, and datasets not yet part of our reference set are not considered.

The results of the second phase presented during the RCC workshop³⁰ show that our approach performs well compared to the other finalist teams, with the highest precision (52.2%; second, 47.0%) and second-best recall (20.5%; best, 34.8%). With respect to F_1 , our approach provides the second-best-performing system for this task (29.5%; best, 40.0%). The results on the manually created holdout set underline that our system performs better with respect to precision compared to the other finalist teams. Given that our models are supervised through a corpus of social science publications, we anticipate limited generalizability across other disciplines and plan to investigate this aspect as part of future work. In this context, the focus of our training data on 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 such as text corpora in the NLP community. In general, however, our approach to using a weakly labelled corpus created from a list of dataset names could be applied in other research domains.

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.

Research Field Classification

Our supervised machine learning approach to the research field classification task performs well on the dataset created from social science publication metadata. A micro F_1 measure of over 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 led to micro F_1 values between 53% and 62%. Considering the evaluation approach, focused on publications from the social sciences, the generalizability across other disciplines remains unclear and requires further research. Even though the classification scheme used may cover neighbouring disciplines (e.g. medicine), the numbers of samples of the training data covering research fields other than the social sciences is limited. Our pragmatic approach of basing our classifications on the abstracts of the publications makes it applicable even in scenarios where the full text of publications is not accessible.

Conclusion

This chapter has provided an overview of our solutions submitted to the Rich Context Competition 2018. With the aim of improving search, discovery and interpretability of scholarly resources, we address three distinct tasks all concerned with extracting structured information about research resources from scientific 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 the 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 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 an important activity for future work, and is a prerequisite for significantly advancing state-of-the-art methods.

Acknowledgements

This work has been partially funded by the 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, 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 J. Chen, M. A. Gonçalves, J. M. Allen, et al. (eds), *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. New York: ACM, pp. 203–212.
- Backes, T. (2018b) The impact of name-matching and blocking on author disambiguation. In A. Cuzzocrea, J. Allan, N. W. Paton, et al. (eds), *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York: ACM, pp. 803–812. Available at: <http://dblp.uni-trier.de/db/conf/cikm/cikm2018.html#Backes18>.
- Bird, S., Dale, R., Dorr, B. J. 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 '08). Marrakech: . European Language Resources Association..
- 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*. Paris: CEUR-WS, pp. 15–27.
- Boland, K., Ritze, D., Eckert, K. and Mathiak, B. (2012) Identifying references to datasets in publications. In P. Zaphiris, G. Buchanan, E. Rasmussen and F. Loizides (eds), *Theory and Practice of Digital Libraries*. Berlin: Springer, pp. 150–161.
- 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*. Montreal: American Society for Information Science, p. 44.
- Finkel, J. R., Grenager, T. and Manning, C. (2005) Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. New Brunswick, NJ: Association for Computational Linguistics, pp. 363–370.
- 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), 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)*. Melbourne: Association for Computational Linguistics 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 M. Bonn, D. Wu, J. S. Downie and A. Martaus (eds), *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries*. Champaign, IL: IEEE Press, pp. 148–157. 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*. Valencia: Association for Computational Linguistics, pp. 427–431.
- 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), 259–289.
- Krämer, T., Klas, C.-P. and Hausstein, B. (2018) A data discovery index for the social sciences. *Scientific Data*, 5, 180064.
- Lewis-Beck, M., Bryman, A. E. and Liao, T. F. (2003) *The Sage Encyclopedia of Social Science Research Methods*. Thousand Oaks, CA: Sage.
- Mikolov, T., Sutskever, I., Chen, K. et al. (2013) Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, pp. 3111–3119.
- Nasar, Z., Jaffry, S. W. and Malik, M. K. (2018) Information extraction from scientific articles: A survey. *Scientometrics*, 117(3), 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 *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*. Portorož, Slovenia: European Language Resources Association, pp. 1862–1868.
- Sahoo, P., Gadiraju, U., Yu, R., Saha, S. and Dietze, S. (2017) Analysing structured scholarly data embedded in web pages. In A. González-Beltrán, F. Osborne, S. Peroni and S. Vahdati (eds), *Semantics, Analytics, Visualization: Enhancing Scholarly Data* (Lecture Notes in Computer Science 9792). Cham: Springer, pp. 90–100.

- Schwartz, A. S. and Hearst, M. A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing, 2003*. River Edge, NJ: World Scientific, 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*, 18(4), 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: Association for Computational Linguistics, pp. 384–394. 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)*. Melbourne: 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>

Notes

- 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 such as *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 5000 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 for it to be considered 10. Word embeddings are clustered using agglomerative clustering with number of clusters set to 500, 600, 700. Word linkage with Euclidean distance is used to minimize the variance within the clusters.
- 25 A glossary of statistical terms as provided at <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 SSOAR metadata can be found at [github/research-field-classifier](https://github.com/research-field-classifier)
- 29 <https://fasttext.cc/>
- 30 The workshop agenda can be found at <https://coleridgeinitiative.org/richcontextcompetition/workshopagenda>. The results of the finalists are presented at <https://youtu.be/PE3nFrEkwoU?t=9865>.

9

FINDING DATASETS IN PUBLICATIONS: THE UNIVERSITY OF PADERBORN APPROACH

Rricha Jalota, Nikit Srivastava,
Daniel Vollmers, René Speck, Michael
Röder, Ricardo Usbeck and Axel-Cyrille
Ngonga Ngomo

Literature Review	130
Project Architecture.....	131
Preprocessing	131
Approach.....	133
Dataset Extraction	135
Evaluation.....	137
Discussion.....	138
Future Agenda.....	139
Appendix.....	140
References.....	140
Notes	141

Literature Review

Previous works on information retrieval from scientific articles are mainly found in the biomedical sciences and computer science, with systems (Westergaard et al., 2018) built using the MEDLINE¹ abstracts, full-text articles from PubMed Central² or the ACL Anthology dataset.³ The documents belonging to the above-mentioned datasets follow a similar format, and thus, several metadata and bibliographical extraction frameworks such as CERMINE (Tkaczyk et al., 2014) 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 such as GROBID (Lopez, 2009), ScienceParse⁴ or ParsCit (Council et al., 2008) did not seem viable options, mainly because these systems were still under development and lacked certain desired features. Hence, building upon the approach of Westergaard et al. (2018), 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 and Kuhn (2017) extracted core claims from scientific articles by first detecting key words 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 and Manning (2011) 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 rank the extracted key phrases, Mahata et al. (2018) devised an approach for processing text documents to train phrase embeddings.

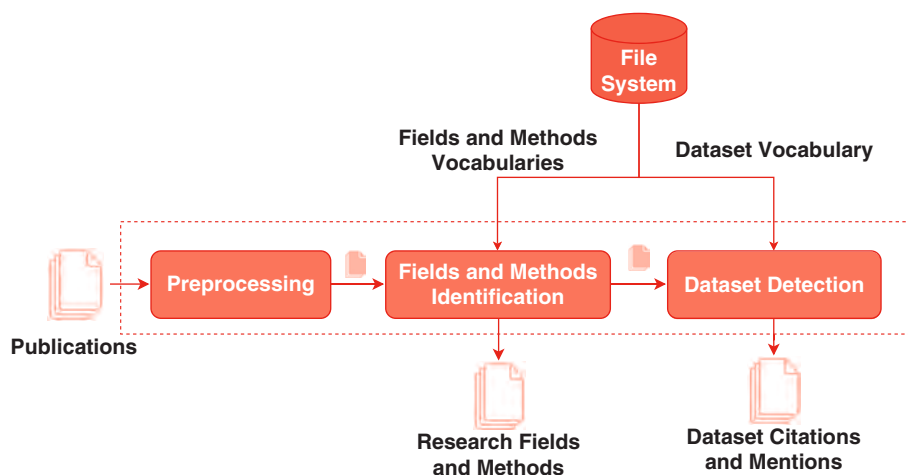


Figure 9.1 Data flow pipeline: the arrows depict the flow of given and generated files between components and the generation of final output files

Not only is the problem of dataset detection and methods and fields identification 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

Our pipeline (shown in Figure 9.1) consisted of three main components: preprocessing; fields and methods identification; and dataset extraction. The preprocessing module read the text from publications and generated some additional files (see below). 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, but 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 Suite⁵ to extract text from PDFs. The `pdf2text` (<https://manpages.debian.org/stretch/poppler-utils/pdf2text.1.en.html>) command served as the first preprocessing step and was called as a subprocess from within a Python script. It was used with the `-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. (2018) for preprocessing. 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, acknowledgements, 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`).⁶

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 headings 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 9.1), 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 in one article. For brevity, we have reported only four main sections in Table 9.1: 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.

Table 9.1 Evaluation of identification of sections in validation data (100 articles)

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` 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. These files were then passed on to the other two components of the pipeline, which are discussed below.

Approach

Research Fields and Methods Identification

Vocabulary Generation and Model Preparation

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, we decided in phase 2 to create our own research methods vocabulary using Wikipedia and DBpedia.⁷ We manually curated a list of all the relevant statistical methods from Wikipedia⁸ 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. (see Table 9.2 for examples).

Table 9.2 Examples from manually curated methods vocabulary

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

Research Fields Vocabulary

For both 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, we additionally created for phase 2, a stopword list of terms that did not contain any domain-specific information (e.g. mixed methods, meta-analysis, narrative analysis).

Word2Vec Model Generation

In this preprocessing 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. The pre-trained Word2Vec model `GoogleNews-vectors-negative300.bin` (Mikolov et al., 2013) was used to extract the vectors of the individual words.

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 naive 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

To find the research fields and methods for a given list of publications, we performed the following steps (first, step 1 was executed for all publications, then steps 2 and 3 were executed iteratively for each publication).

- 1 *Naive research method finder run.* In this step, we executed the naive 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 reran 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 (sectionwise weighting). Here, the noun phrases that came from the methodology section and from the methods listed in JEL classifications (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. We selected the top N ($= 10$) pairs that had the highest weighted cosine similarity. Then 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 postprocessing 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 that 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) are reported in Table 9.3.

Table 9.3 Experimental set-up for training the Word2Vec models for research field (RF) and methods (RM) identification

Computing infrastructure	macOS, 2GHz Intel Core i7 processor, 4 cores, RAM 16GB, 1600 MHz
ATT – RF model	3 min 21 s
ATT – RM model	3 min 19 s
Link to implemented code	https://github.com/nikit91/Jword2vec/tree/rich-context

Dataset Extraction

To identify the datasets in a publication, we followed two approaches and later combined results from both.

In our first approach, *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 in Figure 9.2, 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 such as ‘time’, ‘series’, ‘time series’, and ‘population’. Therefore, we filtered those dataset citations that occurred more often than a certain threshold value ($= 1.20$) multiplied by the median of the frequency distribution and that had fewer than three distinct mentions in a publication. The remaining citations were written to an interim result file.

Table 9.4 depicts the improvement in performance of simple dataset mention search with the inclusion of filtering. The filtering process improved the F_1 measure by 42.86%. Note that, as the validation data consisted of only 100 articles, changing the threshold value to 1.10 or 1.30 did not result in any significant change, hence we have maintained a constant threshold value of 1.20 in our comparison table.

Table 9.4 Performance of simple dataset mention search against validation data

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

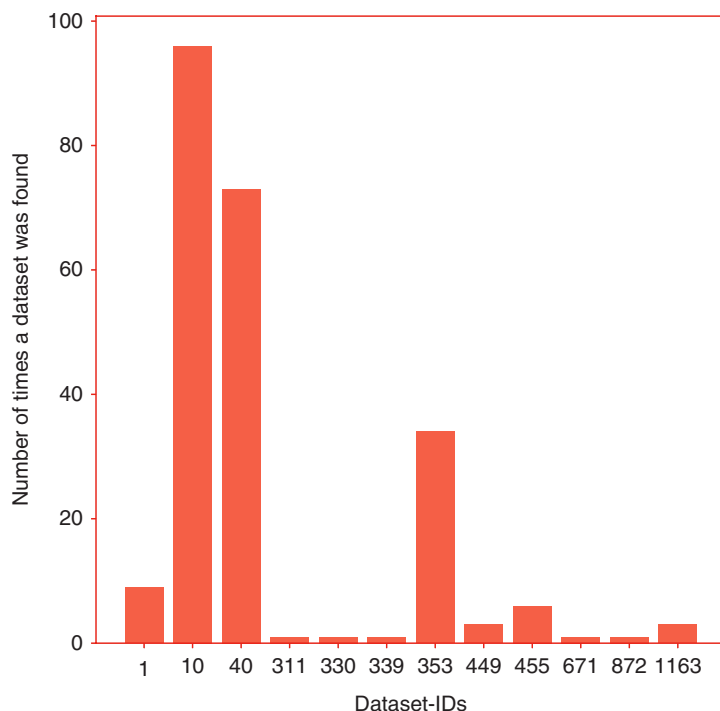


Figure 9.2 Frequency distribution of Dataset Citations

In our second approach, *Rasa-based dataset detection*, we trained an entity extraction model based on conditional random fields (CRFs) using Rasa NLU (Bocklisch et al., 2017). To train the model we used the spaCy Tokenizer⁹ for the preprocessing step. For entity recognition we used BIOES 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 model. In phase 1, the 2500 labelled publications from the training dataset were used for training the Rasa NLU¹⁰ model. Later in phase 2, when the phase 1 holdout corpus was released, we combined its 5000 labelled publications with the previously given 2500 labelled publications and then retrained the model again with these 7500 labelled publications.

The trained CRF-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.

Finally, *we combined 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` occurred more often than a threshold value ($= 1.20$) multiplied by the median of the frequency distribution (the same as for the simple dataset mention search filtering process).

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 of dataset extraction using the evaluation script provided by the competition organizers (see Chapter 5 for more details). This evaluation (see Table 9.5) 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 occurring terms from the Rasa-generated output, based on the frequency distribution of dataset mentions as computed in the simple dataset mention search.

Table 9.5 Quantitative evaluation of datasets against validation data (numbers in parentheses indicate training samples)

Metrics	Phase 1		Phase 2	
	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
F_1	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 the phase 1 holdout corpus.

Tables 9.6 and 9.7 show 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.

Table 9.6 Evaluation of research fields against phase 1 holdout

Pub_id	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

Table 9.7 Evaluation of research methods against phase 1 holdout

Pub_id	Keywords	Phase 1	Phase 2
10328	Thematic content analysis	Thematic analysis	Šidák 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 at all 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 – which contained all data relevant to our work – demanding. As mentioned before, if there was no explicit mention of the key terms such as 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 labelled 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 did not get missed.

Regarding the fields and methods, vocabularies played an immense role in their identification. The vocabularies that were provided by Sage 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 stopwords list of irrelevant terms and looked it up each time, before writing the result to the output file. The goal of stopwords-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 to 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 well as we expected, the 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 stopwords list, just as we did for research field identification, could also have been another workaround.
- 2 For both fields and methods identification, we could also have tried pre-trained embeddings from GloVe¹¹ and fastText.¹²
- 3 As our entity extraction approach to 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 hand-crafted 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 therefore suggest 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 (Jaradeh et al., 2019; Buscaldi et al., 2019). 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 therefore 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 artefacts 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.¹³ The resulting graphs will:

- rely on advanced distributed storage for RDF to scale to the large number of publications available
- be self-feeding, that is, crawl the web for potentially relevant content and make this content available for processing
- be self-repairing, that is, 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 incorrect 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>.

References

- Bocklisch, T., Faulkner, J., Pawlowski, N. and Nichol, A. (2017) Rasa: Open source language understanding and dialogue management. Preprint, arXiv:1712.05181.
- Buscaldi, D., Dessì, D., Motta, E., Osborne, F. and Recupero, D. R. (2019) Mining scholarly data for fine-grained knowledge graph construction. In *Proceedings of the Workshop on Deep Learning for Knowledge Graphs (DL4KG2019) Co-located with the 16th Extended Semantic Web Conference 2019 (ESWC 2019), Portoroz, Slovenia, June 2, 2019*, CEUR-WS. org:Aachen, Germany. pp. 21–30.
- Councill, I. G., Giles, C. L. and Kan, M.-Y. (2008) ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC '08)*. European Language Resources Association, Paris: https://en.wikipedia.org/wiki/European_Language_Resources_Association. pp. 661–667.
- Gupta, S. and Manning, C. (2011) Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pp. 1–9. <https://dblp.uni-trier.de/rec/bibtex/conf/ijcnlp/GuptaM11>
- Jansen, T. and Kuhn, T. (2017) Extracting core claims from scientific articles. In *BNAIC 2016: Artificial Intelligence*. Cham: Springer, pp. 32–46.
- Jaradeh, M. Y., Auer, S., Prinz, M., Kovtun, V., Kismihók, G. and Stocker, M. (2019) Open research knowledge graph: Towards machine actionability in scholarly communication. Preprint, arXiv:1901.10816.
- Lopez, P. (2009) GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In J. L. Borbinha, S. Kapidakis, C. Papatheodorou, G. Tsakonas (eds), *Research and Advanced Technology for Digital Libraries*. Berlin: Springer, pp. 473–474.

- Mahata, D., Kuriakose, J., Shah, R. R. and Zimmermann, R. (2018) Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Stroudsburg PA: Association for Computational Linguistics, pp. 634–639.
- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient estimation of word representations in vector space. Preprint, arXiv:1301.3781.
- Tkaczyk, D., Szostek, P., Dendek, P. J., Fedoryszak, M. and Bolikowski, L. (2014) Cermine: Automatic extraction of metadata and references from scientific literature. In *2014 11th IAPR International Workshop on Document Analysis Systems*. Piscataway, NJ: IEEE, pp. 217–221.
- Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., and Bolikowski, L. (2015) CERMINE: automatic extraction of structured metadata from scientific literature. In *International Journal on Document Analysis & Recognition (IJ DAR)*, vol. 18, no. 4, pp. 317–335, doi: 10.1007/s10032-015-0249-8.
- Westergaard, D., Stærfeldt, H. H., Tønsberg, C., Jensen, L. J. and Brunak, S. (2018) A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Computational Biology*, 14(2).

Notes

- 1 <https://www.nlm.nih.gov/bsd/medline.html>
- 2 <https://www.ncbi.nlm.nih.gov/pmc/>
- 3 <https://www.aclweb.org/anthology/>
- 4 <https://github.com/allenai/science-parse>
- 5 <https://manpages.debian.org/testing/poppler-utils>
- 6 <https://github.com/explosion/spaCy>
- 7 <https://wiki.dbpedia.org/services-resources/ontology>
- 8 https://en.wikipedia.org/wiki/Category:Statistical_methods
- 9 <https://spacy.io/api/tokenizer>
- 10 <https://rasa.com/docs/nlu>
- 11 <https://nlp.stanford.edu/projects/glove>
- 12 <https://fasttext.cc/docs/en/crawl-vectors.html>
- 13 <https://www.w3.org/TR/shacl/>

10

FINDING DATASETS IN PUBLICATIONS: THE SINGAPORE MANAGEMENT UNIVERSITY APPROACH

Philips Kokoh Prasetyo, Amila Silva, Ee-Peng
Lim, and Palakorn Achananuparp

Introduction.....	144
Related Work	144
Data Analysis.....	145
Methods	147
Experiment Results.....	151
Acknowledgements.....	154
References.....	154
Notes	155
Appendix: Technical Documentation	155

Introduction

Automated discovery from scientific research publications is an important task for analysts, researchers and learners as they develop scientific knowledge and use it 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 to 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 dataset discovery from social science research publications, filling the gap for this problem. In this competition, given a corpus of social science research publications, we had to automatically identify datasets used, and then infer the research methods and research fields in the publications. Note that no labelled data were given for research methods and fields identification.

We describe our submission to the first phase of RCC. We performed dataset detection followed by an implicit entity linking approach to tackle the dataset extraction task. We adopted weakly supervised classification for the research method and field identification tasks utilizing the external resource Sage Knowledge as a proxy for weak labels.

Related Work

Extracting information from scientific text has been explored in the past (Peng and McCallum, 2004; Nguyen et al., 2015; Singh et al., 2016). One type of information extraction from scientific articles is extracting keyphrases and relations between them (Augenstein et al., 2017). Luan et al. (2017) propose a semi-supervised sequence tagging approach to extract keyphrases. Augenstein and Søgaard (2017) explore a multi-task deep recurrent neural network approach with several auxiliary tasks to extract keyphrases.

Another type of extraction is citation extraction. Two citation extraction settings have been explored before: reference mining inside the full text (Alves et al, 2018), and citation metadata extraction (Hetzner, 2008; Anzaroot et al., 2014; An et al., 2017). Nasar et al. (2018) have written a survey on information extraction from scientific articles.

Recently, there have been explorations of dataset extraction from scientific text (Boland et al., 2012; Ghavimi et al., 2016a, 2016b). Boland et al. (2012) propose weakly supervised pattern induction to identify references in social science publications. Ghavimi et al. (2016a, 2016b) propose a semi-automatic approach to detecting dataset references in social science texts. Dataset extraction is a challenging task because of the inconsistency and wide range of dataset mention styles in research publications (Ghavimi et al., 2016b).

Data Analysis

The first phase of the RCC competition involved a labelled corpus of 5000 publications as training set, and an additional 100 publications as development set. The RCC organizer kept a separate corpus of 5000 publications for evaluation. Each article in the dataset contained a full-text article and dataset citation labels. The metadata of cited datasets in the corpus were also provided. For research methods and fields, no label information was provided; only Sage social science research method graph and research fields vocabulary was provided. More details on the RCC dataset and competition design can be found in Chapter 5.

Preprocessing

In order to reliably access important structures of paper publications, we parsed all papers using AllenAI Science Parse¹ (Ammar et al., 2018). This parser reads a PDF file and returns title, authors, abstract, sections, and bibliography (references). Since it utilizes machine learning models to parse PDF files, the parsing results may not be 100% accurate. Furthermore, the parser is unable to parse scanned images of old publications. When we were unable to access parsed fields, we fell back on the given text files.

Mention Analysis

There were 5499 and 123 dataset citations in the training and development sets, respectively. Of these citations, 320 in the training set and 6 in the development set did not have mentions information. We analysed the paper sections where the dataset mentions commonly occur. Tables 10.1 and 10.2 show the 12 most common sections mentioning datasets in the training and development sets, respectively. 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 the training set

Section header	Mention frequency
Abstract	2548
Reference titles	1997
Discussion	1390
Results	836
Methods	804
Introduction	530
Statistical analysis	285
Comment	279

(Continued)

Table 10.1 (Continued)

Section header	Mention frequency
Acknowledgements	261
Materials and methods	254
Study population	227
Data	214

Table 10.2 Top 12 sections mentioning datasets in the 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
Data	5

Citation Analysis

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

Table 10.3 Citation network statistics

Number of nodes	5000
Number of edges	1222
Network density	0.0098%

Initially, we proposed an approach utilizing a citation network based on an intuition that datasets, research methods, and research fields are shared by: the same or similar issues; the same or similar context; the same or similar authors and communities; the same or similar metrics used in the publication. However, based on Table 10.3, we learned that exploring rich context using a citation network is not viable at this stage because most papers listed in publications’ bibliographies are not available in the train-

ing set, and therefore, the citation network becomes very sparse with much unknown information. Figure 10.1 visualizes the citation network. As we can see, most papers only have one edge, although the average number of papers in the bibliography list is 34.5. Furthermore, only 1466 out of 5000 publications are not isolated (having at least one bibliography paper in training data). For this reason, we drop our idea of utilizing a citation network at this stage. Nevertheless, we believe that bibliographies contain important signals and information about datasets and research fields.

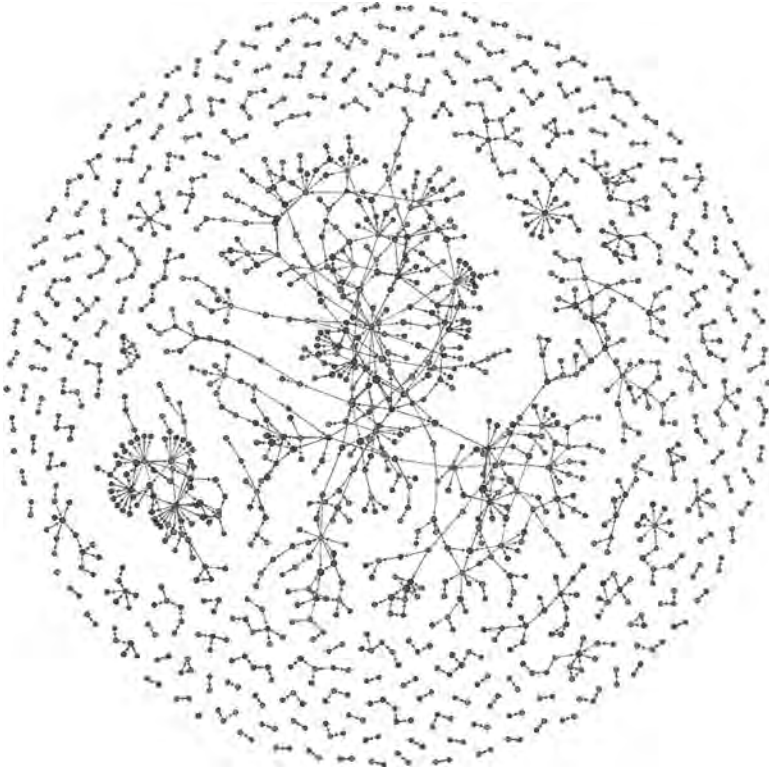


Figure 10.1 Citation network from training data. Green nodes are publications with dataset citations, and red nodes are publication without dataset citations. Isolated nodes are not visualized.

Methods

In this section we describe our approach to the RCC tasks of dataset extraction, research methods identification, and research fields identification.

Dataset Extraction

We employed 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 mined dataset mentions for the remaining publications in the dataset recognition subtask.

For dataset detection, we utilized paper titles in bibliographies (reference lists) combined with explicit research methods mentions to detect whether a publication cites a dataset or not. Explicit research methods mentions were determined based on an exact match between paper title and Sage research methods vocabulary. We trained a support vector machine (SVM) classifier using explicit research method mentions and n -gram features from paper titles in bibliographies. We used the SVM classifier to classify each publication: if the classifier gave a positive label, then we proceeded to the dataset recognition subtask, otherwise we ignored the publication.

For dataset recognition, we used an implicit entity linking approach. We started with the naive Bayes model, which can be regarded as a standard information retrieval baseline, and an entity indicative weighting strategy was used to improve the model. In order to calculate the word distribution of each dataset, we represented 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 Ghavimi et al. (2016b). All these texts related to a particular dataset are considered as a single string, and we calculated the word distribution as follows. Let 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 $w \in W$ is

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

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 derived from the count of w occurrences in the text extracted for each dataset. In order to give more priority to dataset-indicative words, we improved the final objective function of our model as follows:

$$\ln(p(d | w)) \propto \sum_{w \in W} \beta(w) \ln(p(w | d)),$$

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

$$\beta(w) = \ln \left(1 + \frac{E}{df(w)} \right),$$

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 takes approximately 5 minutes.

Then, for a given unseen publication, we used the same rule-based approach (Ghavimi et al., 2016b) to filter a few relevant sentences, and datasets were ranked by $\ln(p(d|w))$ to select the most suitable datasets. In order to select exact datasets related to a particular publication, we selected the top 10 datasets ranked using the above approach. The confidence probability related to the top 10 datasets was then normalized and the datasets with normalized probability higher than a predefined threshold value were selected. We returned the entity-indicative words as relevant dataset mentions.

Research Methods Identification

Since we did not have labelled training data for this task, we used explicit research method mentions (based on exact matches 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 generated a training set for research method classification utilizing sentences that explicitly mentioned research methods in a publication. Publication title and the sentences mentioning research methods served as context information for a specific research method. In order to reduce noisy weak signals, we applied a minimum support of three sentences in a publication. We excluded research methods only mentioned once or twice in a publication. We also excluded research methods only mentioned in fewer than 10 different publications from the training set. Finally, we had 133 research methods with sufficient context information as our training data. This number is 20.18% of the 659 research methods in Sage research method graph.

We used the training data to train a logistic regression classifier to classify research methods from publication titles and sentences. We utilized n -gram features from publication titles and sentences for the classifier. We applied the logistic regression classifier to recommend the top three research methods based on logistic regression probability scores.

This approach can be extended by utilizing research method graphs to expand the context. Context information comes not only from sentences in publications, but also 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

Similarly to research methods identification, this task did not have labelled training data. We only had access to a list of Sage research fields. These are organized hierarchically into three levels (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 crawled the top search results from Sage Knowledge.² From the search result snippets, we collected information such as title and abstract on various publications including cases, major works, books, handbooks, and dictionaries. We excluded videos and encyclopedias. Due to the sparseness of Sage Knowledge, we excluded all research fields with less than 10 search results. In the end, we had samples of 414 L3 research fields, under 101 L2 research fields, and 10 L1 research fields, covering 20.87% of the 1984 L3 research fields, and 67.79% of the 149 L2 research fields in the list of Sage research fields. We used this data to train research fields classifiers.

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

- 1 Get the top five L3 research fields, the top four L2 research fields, and the top three L1 research fields.
- 2 Assign an initial score for each research field based on its ranking:

$$v(f_i) = \frac{K-i}{K},$$

where K is the length of the top k , and i is the ranking of a research field f . For example, the top five L3 research fields have initial score [1, 0.8, 0.6, 0.4, 0.2], the top four L2 have [1, 0.75, 0.5, 0.25], and the top three L1 have [1, 0.666, 0.333].

- 3 Update the score by multiplying each score by the scores of matching research fields at upper level, and by 0 otherwise:

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

where L is the level of research field f and its upper levels. Here are examples of score updates:

- 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 $0.8 \times 0.5 \times 1 = 0.4$.
 - 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 After collecting all scores from L2 and L3, if we see 'Soc-2', 'Soc-2-4' and 'Soc-2-8', we can exclude 'Soc-2' because 'Soc-2-4' and 'Soc-2-8' are parts of 'Soc-2'.

- 5 Rerank L2 and L3 research fields based on the score.
- 6 Return research fields having a score of at least 0.4.

We also built three other naïve Bayes classifiers for L1, L2, and L3 using paper title in bibliography section to exploit more context from the list of papers in bibliography section. We believe that a publication from a certain field also cites other publications from the 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 the top research fields from the bibliography. Finally, we combine the top research fields from paper titles and abstracts with results from the bibliography.

Experiment Results

We discuss our experiment results for each task in this section. We use standard precision, recall, and F_1 score as evaluation metrics.

Dataset Extraction

First, we analyse our experiment for the dataset detection subtask comparing the naïve Bayes and SVM classifiers. Using only paper titles in bibliographies and explicit research method mentions, the naïve Bayes and SVM classifiers are able to achieve F_1 scores of 0.88 and 0.92, respectively. Since the SVM outperforms naïve Bayes, we use SVM for our dataset detection module. Table 10.4 shows detailed dataset detection results on the development set.

Table 10.4 Dataset detection results on the development set

Classifier	Precision	Recall	F_1
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 the development set. Table 10.5 summarizes the results. As shown in the table, performing dataset detection before extraction significantly improves the dataset extraction on the development set.

Table 10.5 Dataset extraction results on the development set

Method	Precision	Recall	F_1
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 the test set

Dataset	Precision	Recall	F_1
Test set (phase 1)	0.17	0.10	0.13

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

Research Methods Identification

We only consider naive Bayes and logistic regression classifiers for research method identification because they naturally output probability scores. We perform 5-fold cross validation to evaluate classification performance, and the result can be seen in Table 10.7. The logistic regression classifier outperforms naive Bayes with an F_1 score of 0.86 in classifying 133 research methods.

Table 10.7 F_1 score for research method classification

Classifier	F_1
Naive Bayes	0.55
Logistic regression	0.86

Research Fields Identification

We perform 5-fold cross-validation to evaluate our classification of L1, L2 and L3 research fields. Table 10.8 shows the results using n -gram features from paper title and abstract, while 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 a large number of research field labels. We decided to use SVM for research field identification at the publication level because SVM is generally better than naive Bayes. On the other hand, we decided to use naive Bayes for research field identification at the bibliography level because naive Bayes prefers to have more accurate L2 and L3 research fields.

Table 10.8 F_1 score for research field classification at 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 F_1 score for research field classification at 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

Lessons Learned

Extraction of research datasets, associated research methods and fields from social science publication is challenging, yet an important problem for organizing social science publications. We have described our approach to the RCC challenge, and Table 10.10 summarizes our approach. Besides publication content such as paper titles, abstract and full text, our approach also leverages on bibliography information. 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 the F_1 score on 5-fold cross-validation, we have no good way to evaluate research method and research field identification without ground truth labels.

Our methods are unable to automatically extract and recognize new datasets, research methods and fields. An extension to automatically handle such cases using an advanced 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 in Chapters 6–9.

Our model did not perform well on the test set, and we were unable to advance to the second phase. Nevertheless, we recommend our approach as a baseline method as it is simple, efficient, and fast to train. From this competition, we have learned that a lack of labelled training data is a huge challenge, and it directs us to other external resources (i.e. Sage Knowledge) as proxy labels. 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 bibliographies, lack of access to these publication makes it difficult to exploit the citation network. Expanding labelled data from the list of papers in bibliographies will be very beneficial to improve the rich context of paper publications and datasets.

Acknowledgements

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

References

- Alves, D. R., Colavizza, G. and Kaplan, F. (2018) Deep reference mining from scholarly literature in the arts and humanities. *Frontiers in Research Metrics and Analytics*.
- Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T., H.-H., Peters, M. E., Power, J., Skjonsberg, S., Wang, L. L., Wilhelm, C., Yuan, Z., van Zuylen, M. and Etzioni, O. (2018) Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 84–91.
- An, D., Gao, L., Jiang, Z., Liu, R. and Tang, Z. (2017): Citation metadata extraction via deep neural network-based segment sequence labeling. In *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. New York: ACM, pp. 1967–1970.
- Anzaroot, S., Passos, A., Belanger, D. and McCallum, A. (2014) Learning soft linear constraints with application to citation field extraction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 593–602.
- Augenstein, I., Das, M., Riedel, S., Vikraman, L. and McCallum, A. (2017) Semeval 2017 task 10: ScienceIE – extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 546–555.

- Augenstein, I. and Søgaard, A. (2017) Multi-task learning of keyphrase boundary classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Stroudsburg, PA: Association for Computational Linguistics, pp. 341–346
- Boland, K., Ritze, D., Eckert, K. and Mathiak, B. (2012) Identifying references to datasets in publications. In P. Zaphiris, G. Buchanan, E. Rasmussen and F. Loizides (eds), *Theory and Practice of Digital Libraries*. Berlin: Springer, pp. 150–161.
- Ghavimi, B., Mayr, P., Lange, C. Vahdati, S. and Auer, S. (2016a) A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use*, 36(3–4), 171–187.
- Ghavimi, B., Mayr, P., Vahdati, S. and Lange, C. (2016b) Identifying and improving dataset references in social sciences full texts. In *ELPUB 2016: the 20th International Conference on Electronic Publishing Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Amsterdam: IOS Press, pp. 171–187.
- Hetzner, E. (2008) A simple method for citation metadata extraction using hidden Markov models. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM, pp. 280–284.
- Luan, Y., Ostendorf, M. and Hajishirzi, H. (2017) Scientific information extraction with semi-supervised neural tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, pp. 2641–2651. <https://www.aclweb.org/anthology/D17-1279/>
- Nasar, Z., Jaffry, S. W. and Malik, M. K. (2018) Information extraction from scientific articles: a survey. *Scientometrics*, 117, 1931–1990.
- Nguyen, V. C., Chandrasekaran, M. K., Kan, M.-Y. and Lee, W. S. (2015) Scholarly document information extraction using extensible features for efficient higher order semi-CRFs. In *JCDL '15: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: ACM, pp. 61–64.
- Peng, F. and McCallum, A. (2004) Accurate information extraction from research papers using conditional random fields. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 329–336.
- Singh, M., Barua, N., Palod, P., Garg, M., Satapathy, S., Bushi, S., Ayush, K., Rohith, K. S., Gamidi, T., Goyal, P. and Mukherjee, A. (2016) Ocr++: A robust framework for information extraction from scholarly articles. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Stroudsburg, PA: Association for Computational Linguistics, pp. 3390–3400.

Notes

- 1 <https://github.com/allenai/science-parse>
- 2 <http://sk.sagepub.com/browse/>

Appendix: Technical Documentation

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

11

FINDING DATASETS IN PUBLICATIONS: THE SYRACUSE UNIVERSITY APPROACH

Tong Zeng and Daniel Acuna

Introduction.....	158
The Dataset.....	158
The Proposed Method.....	159
Results.....	162
Conclusion.....	163
Acknowledgements.....	164
References.....	164
Notes	165

Introduction

Science is fundamentally an incremental discipline that depends on previous scientists' work. Datasets form an integral part of this process and therefore should be shared and cited like any other scientific output. This ideal is far from reality: the credit that datasets currently receive does not correspond to their actual usage (Zeng et al., 2020). 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 to close this gap by inviting scientists to produce automated dataset mention and linkage detection algorithms. In this chapter, 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. Ghavimi et al. (2016, 2017) use a relatively simple TF-IDF representation with cosine similarity for matching dataset identification in social science articles. Their method consists of three 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. Singhal and Srivastava (2013) 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 computationally 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 Lu et al. (2012). 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 chapter, 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 (bi-LSTM) sequence to sequence model paired with a conditional random field (CRF) inference mechanism. The architecture is similar to that of 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 achieved 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.¹ The challenge comprised several phases, and participants moved through the phases depending on their performance. We only analyse data from the first phase. This

phase contained a list of datasets and a labelled corpus of around 5000 publications. Each publication was labelled indicating whether a dataset was mentioned within it and which part of the text mentioned it. The challenge used an accuracy measure for measuring the performance of the competitors and also the quality of the code, documentation, and efficiency.

We adopted the CoNLL 2003 format (Tjong Kim Sang and De Meulder, 2003) to annotate whether a token is a part of dataset mention. Concretely, we use the tag DS to denote a dataset mention; a B- prefix indicates that the token is the beginning of a dataset mention, an I- prefix indicates that the token is inside a dataset mention, and O denotes a token that is not a part of a dataset mention. We put each token and its tag (separated by a horizontal tab control character) on one line, and use the end-of-line (\n) control character as separator between sentences (see Table 11.1). The dataset was randomly split by 70%, 15%, 15% for training set, validation set and testing set, respectively.

Table 11.1 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 bi-LSTM 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 constraints between label transitions (e.g., mention-to-no-mention-to-mention has low probability). While this is a standard architecture for modelling sequence labelling, the application to our particular dataset and problem is new.

We now describe in more detail the choices of word representation, hyperparameters and training parameters. A schematic view of the model is given in Figure 11.1 and the components are as follow.

- 1 Character encoder layer: treat a token as a sequence of characters and encode the characters by using a bi-LSTM to get a vector representation.
- 2 Word embedding layer: mapping each token into fixed-size vector representation by using a pre-trained word vector.
- 3 Bi-LSTM layer: make use of bi-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 distribution of labels.
- 5 CRF layer: find the most likely sequence of labels.

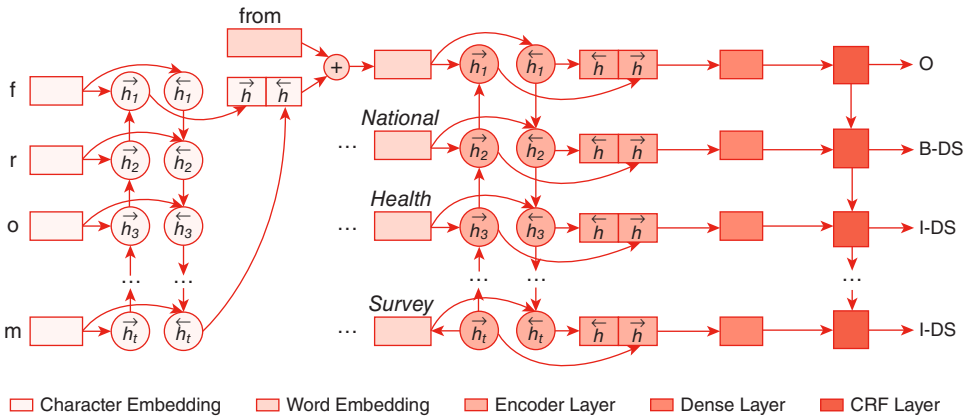


Figure 11.1 Network architecture of bi-LSTM CRF network

Character Encoder

Similar to the bag of words assumption, a word could be composed of characters sampled from a bag of characters. Previous research (Santos and Zadrozny, 2014; Jozefowicz et al., 2016) 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 bi-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. Specifically, the word embedding layer contains a word embedding matrix $M^{\text{tkn}} \in \mathbb{R}^{|V| \times d}$, where V is the vocabulary of the tokens and d is the size of the embedding vector. The embedding matrix was initialized by a pre-trained GloVe vectors (Pennington et al., 2014), 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 look up the embedding matrix to get its vector representation. Finally, we get a sequence of vectors as input for the encoder layer.

LSTM

A recurrent neural network (RNN) is a type of artificial neural network which recurrently takes the output of the previous step as input of the current step. This recurrent nature allows it to learn from sequential data, for example, the text which consists of a sequence of words. An RNN could in theory capture contextual information in variable-length sequences, but it suffers from gradient exploding/vanishing problems (Pascanu et al., 2013). The long short-term memory (LSTM) architecture was proposed by Hochreiter and Schmidhuber (1997) to cope with these gradient problems. Similar to a standard RNN, the LSTM network also has a repeating module called an 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 an output gate controlling the information flowing out of the cell. The LSTM could be defined formally by the following equations:

$$\begin{aligned} i_t &= \sigma(W_i x_t + W_i h_{t-1} + b_i), \\ f_t &= \sigma(W_f x_t + W_f h_{t-1} + b_f), \\ g_t &= \tanh(W_g x_t + W_g h_{t-1} + b_g), \\ o_t &= \sigma(W_o x_t + W_o h_{t-1} + b_o), \\ c_t &= f_t \otimes c_{t-1} + i_t \otimes g_t, \\ h_t &= o_t \otimes \tanh(c_t), \end{aligned}$$

where x_t is the input at time t , W is the weights, b is the bias. The σ is the sigmoid function, \otimes denotes the dot product, c_t is the LSTM cell state at time t and h_t is hidden state at time t . i_t , f_t , o_t and g_t are referred to 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 bi-LSTM (Graves et al., 2013) satisfies this information need by

using two LSTMs. Specifically, 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:

$$h_t = [\overrightarrow{h}_t \quad \overleftarrow{h}_t].$$

Finally, the outcomes of the states are taken by a CRF layer (Lafferty et al., 2001) that takes into account the transition nature of the beginning, intermediate, and end of mentions.

Results

In this work we wanted to propose a model for the Rich Context Competition challenge. We propose a relatively standard architecture based on the bi-LSTM CRF network. We now describe the evaluation metrics, hyperparameter 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 labelling tasks. It varies from 0 to 1, the higher the better. Our method achieved a relatively high F_1 of 0.885 for detecting mentions.

Table 11.2 Model search space and best assignments

Hyperparameter	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 use the Adam optimizer with learning rate 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 overfitting, we use L2 regularization set to 0.01 and we also use dropout rate equal to 0.5. We trained eight models with a combination of different GloVe vector size (50, 100, 300 and 300) and dropout rate (0.0, 0.5). The hyperparameter settings are shown in Table 11.2.

Table 11.3 Performance of proposed network

Model	GloVe size	Dropout rate	Precision	Recall	F_1
1	50	0.0	0.884	0.873	0.878
2	50	0.5	0.877	0.888	0.882
3	100	0.0	0.882	0.871	0.876
4	100	0.5	0.885	0.885	0.885
5	200	0.0	0.882	0.884	0.883
6	200	0.5	0.885	0.880	0.882
7	300	0.0	0.868	0.886	0.877
8	300	0.5	0.876	0.878	0.877

The test performances are reported in Table 11.3. The best model is trained by word vector size 100 and dropout rate 0.5, with F_1 score 0.885 (Table 11.3), 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 128GB 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 is that 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 linked 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 bi-LSTM 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 dataset noise is 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, as in this case. These benefits could also translate to better linkage. We are exploring new architectures using self-attention and multiple-head attention. We hope to share these approaches in the near future.

There are a number of improvements that we could make in the future. A first improvement would be to use non-recurrent neural architectures such as the Transformer which has been shown to be faster and a more effective learner than RNNs. 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* (Zeng et al., 2020) – in contrast to the most common types of citations to publications. The location of such 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 chapter, 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 strive to improve the quality of these datasets so that techniques like this one can be more broadly applied.

Acknowledgements

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

References

- Ghavimi, B., Mayr, P., Lange, C., Vahdati, S. and Auer, S. (2017) A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use*, 36(3–4), 171–187.
- Ghavimi, B., Mayr, P., Vahdati, S. and Lange, C. (2016) Identifying and improving dataset references in social sciences full texts. Preprint, arXiv:1603.01774 [cs].
- Graves, A., Mohamed, A. and Hinton, G. (2013) Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. Piscataway, NJ: IEEE, pp. 6645–6649.
- Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N. and Wu, Y. (2016) Exploring the limits of language modeling. Preprint, arXiv:1602.02410.
- Lafferty, J., McCallum, A. and Pereira, F. C. N. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco: Morgan Kaufmann, pp. 282–289.

- Lu, M., Bangalore, S., Cormode, G., Hadjieleftheriou, M. and Srivastava, D. (2012) A dataset search engine for the Research Document Corpus. In *2012 IEEE 28th International Conference on Data Engineering*. Piscataway, NJ: IEEE, pp. 1237–1240.
- Pascanu, R., Mikolov, T. and Bengio, Y. (2013) On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, Stroudsburg, PA: International Machine Learning Society. pp. 1310–1318.
- Pennington, J., Socher, R. and Manning, C. (2014) GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Red Hook, NY: Curran, pp. 1532–1543.
- Santos, C. D. and Zadrozny, B. (2014) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. Red Hook, NY: Curran, pp. 1818–1826.
- Singhal, A. and Srivastava, J. (2013) Data Extract: Mining context from the Web for dataset extraction. *International Journal of Machine Learning and Computing*, 3(2), 219–223.
- Tjong Kim Sang, E. F and De Meulder, F. (2003) Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, Volume 4. Morristown, NJ: Association for Computational Linguistics, pp. 142–147.
- Zeng, T., Wu, L., Bratt, S. and Acuna, D. E. (2020) Assigning credit to scientific datasets using article citation networks. *Journal of Informetrics*, 14(2), 101013.

Notes

- 1 <https://coleridgeinitiative.org/richcontextcompetition>

12

THE FUTURE OF AI IN RICH CONTEXT

Paco Nathan

The Case For Rich Context.....	169
Use Cases For Rich Context	171
System Architecture Overview.....	176
Trends, From Origins To Near-Term Future Projections.....	178
Summary.....	178
References.....	179
Notes	179

The setting for rich context originated in needs to analyse confidential microdata for evidence-based policy-making.¹ The nature of that work involves collaboration using *linked data*,² 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 sensitive data within their secure environments, and otherwise leave the process unexamined.

However, there is 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 have 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, and so on, 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 behaviour, 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 microdata and its related *metadata*³ (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*.⁴ Other datasets may require data use authorizations 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, and so on:

- *Collaboration and workspace*, where researchers collaborate within a secured environment, having obtained authorizations via non-disclosure agreements, data use authorizations, and so on
- *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 data⁵ required for AI applications in industry, as described Andrew Ng.

In general, guidance for rich context can be drawn from the FAIR (Wilkinson et al., 2016) 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*⁶ to represent metadata about datasets, researchers, projects, agencies, and so on (including the computational agents involved) as distinct entities connected through relations that model their linkage.

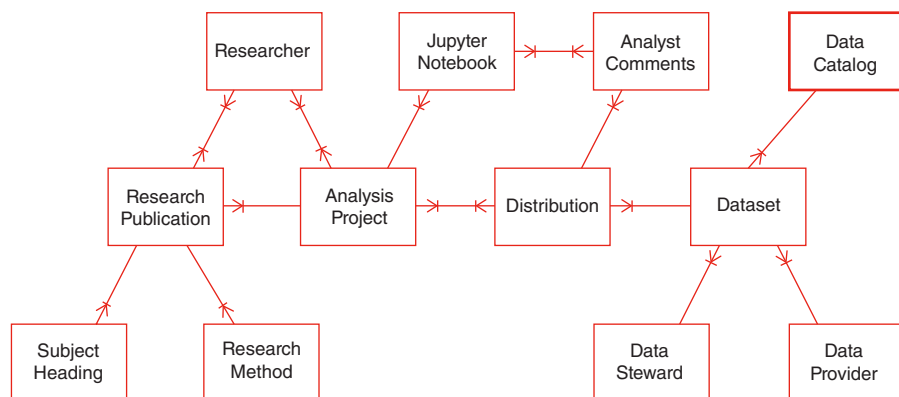


Figure 12.1 Graph relations example

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 Competition⁷ held from September 2018 to 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*,⁸ 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 analyse, 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, and so on, 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 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. There is clearly a need for search and discovery, although other use cases can help improve the discovery process and enhance social science research. The following subsections 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 neighbouring 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 is generally more effective for user experience 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*:⁹ 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*¹⁰ used in the competition need to be generalized and extended, then used to analyse the ongoing stream of published social science research. This work provides potential benefits for the publishers, for example helping them analyse 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 labour-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*,¹¹ which can help add useful annotations to the graph, as shown in the following diagram:

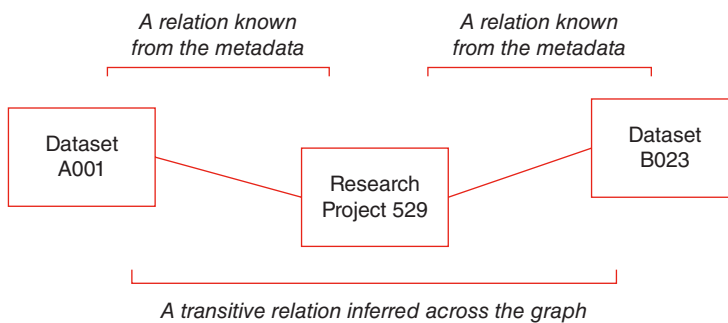


Figure 12.2 Transitive relations

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*¹² 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,¹³ Semantic Scholar,¹⁴ 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 analysed 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 analyse:

- 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 is now captured for others to study, use for training new staff, and so on.

Note that the ADRF framework must provide means for customizing and configuring these kinds of axioms, so that data stewardship 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 ADRF framework represents a *data catalogue* 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 catalogues and datasets.

For example, the W3C Data Activity coordinates a wide range of metadata standards, including:

- DCAT¹⁵ – metadata about data catalogues
- VoID¹⁶ – metadata about datasets
- DCMI¹⁷ – Dublin Core metadata terms
- SKOS¹⁸ – ‘simple knowledge organization system’.

These represent controlled vocabularies¹⁹ described in OWL²⁰ and based atop RDF.²¹ 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 catalogue*.²² Taken together, localized extensions of these open standards represent an ontology²³ – 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*²⁴ 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 complexities²⁵ of data governance in industry data science work. For instance:

- Amundsen²⁶ from Lyft
- Marquez²⁷ from WeWork
- WhereHows²⁸ from LinkedIn
- Databook²⁹ 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. TopQuadrant³⁰ and AstraZeneca³¹ 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 ODPi³² open standard Egeria³³ and related work by Chessell (2016), including the Apache Atlas³⁴ 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.³⁵ A new rich context feature set is being added to JupyterLab,³⁶ 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,³⁷ metadata explorer,³⁸ 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*,³⁹ based on layered structures of artificial neural networks. In particular, *graph embedding* (see Tong, 2019) 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, and so on. 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 with earlier and generally less effective approaches which relied on graph queries applied to the full data. Note that the winning team⁴⁰ in the Rich Context Competition was from Allen AI,⁴¹ 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 PyTorch⁴² (from Facebook) and the more recent Ray⁴³ (from UC Berkeley RISElab).

System Architecture Overview

Figure 12.3 illustrates a proposed system architecture for rich context as an additional module in the ADRF framework. 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.

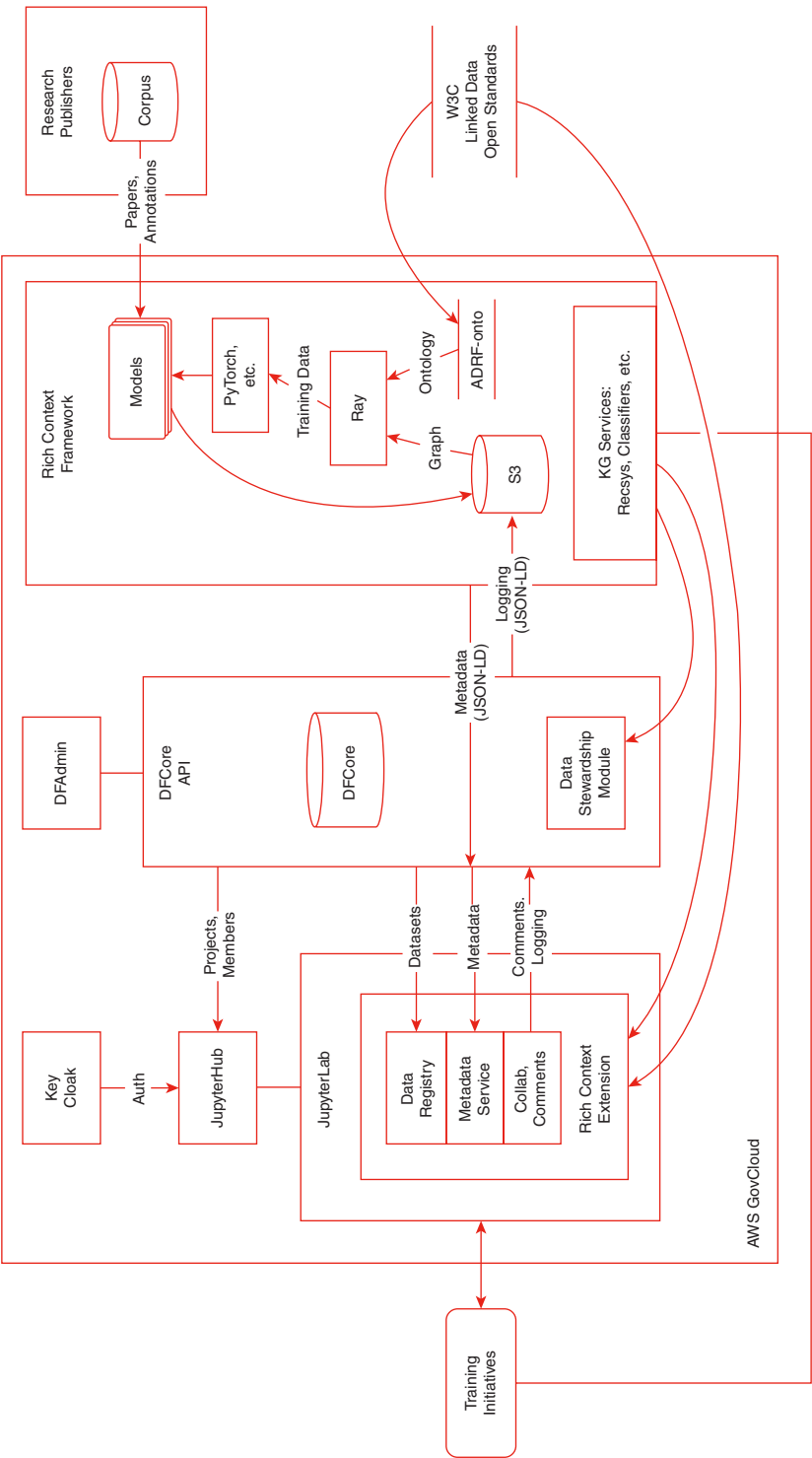


Figure 12.3 Rich context module

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 analyse and report about sensitive data, while maintaining security and privacy compliance. That is 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 is 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 is 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 microdata 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 e-commerce platforms and research in life sciences, social science research presents nuances and new challenges that have not previously been addressed.

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 catalogue. 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 the latter step, in the broader sense rich context helps establish a community focused on contributing code plus knowledge into the research process.

References

- Chessell, M. (2016) The case for open metadata. *Frontiers in Data Science*.
- Hellerstein, J. M., Sreekanti, V., Gonzalez, J. E., Dalton, J., Dey, A., Nag, S. et al. (2017) Ground: A data context service.
- Lin, X. V., Socher, R. and Xiong, C. (2018) Multi-hop knowledge graph reasoning with reward shaping. Preprint, arXiv:1808.10568.
- Tong, F. (2019) Graph embedding for deep learning.
- 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

Notes

- 1 https://en.wikipedia.org/wiki/Foundations_for_Evidence-Based_Policymaking_Act
- 2 <https://www.w3.org/DesignIssues/LinkedData>
- 3 <https://en.wikipedia.org/wiki/Metadata>
- 4 <https://www.data.gov/>
- 5 <https://www.youtube.com/watch?v=21EiKfQYZXc&feature=youtu.be>
- 6 https://en.wikipedia.org/wiki/Knowledge_base
- 7 <https://coleridgeinitiative.org/richcontextcompetition/#problemdescription>
- 8 https://en.wikipedia.org/wiki/Entity_linking
- 9 For a sample of recent research papers regarding link prediction through graph embedding, see these arXiv results: <https://arxiv.org/search/?query=%22link+prediction%22+%22graph+embedding%22&searchtype=all&abstracts=show&order=&size=50>.
- 10 One of the better resources online for entity linking is NLP-progress which specifically tracks the state-of-the-art papers, along with their scores on recognized benchmarks. See http://nlpprogress.com/english/entity_linking.html
- 11 https://en.wikipedia.org/wiki/Transitive_relation

- 12 For a comparison of approaches based on reinforcement learning and embedding, see Lin et al. (2018). See <https://arxiv.org/abs/1808.10568>
- 13 <https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema>
- 14 <http://s2-public-api-prod.us-west-2.elasticbeanstalk.com/corpus/>
- 15 <https://www.w3.org/TR/vocab-dcat/>
- 16 <https://www.w3.org/TR/void/#dataset>
- 17 <https://dublincore.org/specifications/dublin-core/dcmi-terms/>
- 18 <https://www.w3.org/TR/swbp-skos-core-guide/>
- 19 https://en.wikipedia.org/wiki/Controlled_vocabulary
- 20 <https://www.w3.org/OWL/>
- 21 <https://www.w3.org/RDF/>
- 22 <https://www.w3.org/TR/vocab-dcat/#conformance>
- 23 [https://en.wikipedia.org/wiki/Ontology_\(information_science\)](https://en.wikipedia.org/wiki/Ontology_(information_science))
- 24 <https://github.com/Coleridge-Initiative/adrf-onto/wiki>
- 25 A good survey paper about these issues is Hellerstein et al. (2017), based on research by UC Berkeley RISElab. See <https://rise.cs.berkeley.edu/blog/publication/ground-data-context-service-2/>
- 26 <https://eng.lyft.com/amundsen-lyfts-data-discovery-metadata-engine-62d27254fbb9>
- 27 <https://marquezproject.github.io/marquez/>
- 28 <https://github.com/linkedin/datahub>
- 29 <https://eng.uber.com/databook/>
- 30 <https://www.topquadrant.com/>
- 31 <https://www.astrazeneca.com/>
- 32 <https://www.odpi.org/>
- 33 <https://www.odpi.org/projects/egeria>
- 34 <https://atlas.apache.org/#/>
- 35 <https://jupyter.org/>
- 36 <https://jupyterlab.readthedocs.io/en/stable/>
- 37 <https://github.com/jupyterlab/jupyterlab-data-explorer>
- 38 https://github.com/jupyterlab/jupyterlab-metadata-service/blob/master/press_release.md
- 39 https://en.wikipedia.org/wiki/Deep_learning
- 40 <https://ocean.sagepub.com/blog/an-interview-with-the-allen-institute-for-artificial-intelligence>
- 41 <https://allenai.org/>
- 42 <https://pytorch.org/>
- 43 <https://docs.ray.io/en/latest/>

INDEX

- ACL Anthology Reference Corpus, 111
- added value, 16–18
- administrative records, 40, 46
- ADRF framework, 173–8
- advisory services, 15
- algorithms, 20, 23, 25, 32, 139, 158
- Allen, Paul G., 84
- Allen, R.B., 6, 49
- Allen AI (company), 4, 7, 176
 - Science Parse, 145
- Allen Institute for Artificial Intelligence, 84, 90–1
- altmetrics, 33–4
- Amundsen discovery and metadata engine, 47
- analysis platforms, 47
- anonymization, 14, 32
- artificial intelligence (AI), 4,
 - 84, 169–70, 178–9
 - human-in-the-loop* type, 8, 170, 179
- AstraZeneca, 175
- Augenstein, I., 144
- automation of the search for datasets, 2–3,
 - 12, 54, 144, 158, 170
- backward induction, 23
- badging of data, 33–6
- Ball, A., 19
- Basic Formal Ontology (BFO), 43
- bidirectional long short-term memory (bi-LSTM), 158, 160–3
- ‘big data’, 28–9, 168
- blockchains, 46
- Boland, K., 18, 144
- Bush, Vannevar, 2
- Census Bureau, US, 3
- CERN Open Data Portal, 28
- challenges of our time, 168
- Chessell, M., 175
- CISER Data Archive, 28
- citations, 18, 21–3, 45, 144–8, 158
 - types of, 164
- classifiers, 148, 173
- cloud computing, 6, 49
- Coleridge Initiative, 103, 144, 158
- communities of practice, 78
- conditional random field (CRF) models,
 - 117, 123, 136, 158–63
- confidentiality, 21, 168
- Consortium of European Social Science Data Archives, 44
- content analysis, 6
- contextualization, 30–7

- Controlled Lots of Copies Keep Stuff Safe (CLOCKSS), 46
- 'created datasets', 19–21, 24–5
- Creative Commons Licenses, 46
- creativity of research teams, 22–3
- CRedit, 32
- CrossRef, 45
- cultural infrastructure, 5, 28, 31–2
- curation of datasets, axioms for, 174
- data
 - access to, 2, 17, 32, 169
 - commercial, 31–2
 - culture surrounding, 28
 - importance of, 28
 - norms relating to, 29, 31
 - as an output in itself, 29–30
 - reproduction of, 30
 - for research purposes, 29
 - technological infrastructure for, 30–1
 - tools for handling of, 29
 - see also* 'big data'
- data cubes, 6, 47–8
- data elements, 40
- data mining, 33
- data processing, speed of, 28
- 'data shopping', 17
- databases, searching of, 30
- DataCite, 18, 25, 30, 41, 45
- dataset impact factors, 16, 19, 24
- dataset maps, 22
- dataset mentions in publications, 19, 21, 24, 145–6, 159
- dataset metadata, 168
- dataset schemas, 41–2
- dataset similarity indices, 22
- datasets
 - categorization of, 19–20
 - classification of, 30
 - definitions of, 5, 19, 40
 - extracted, 22
 - joint usability of, 16
 - number used in a single paper, 20, 24
 - text associated with, 45
 - usage of, 5, 12, 24–5, 32, 47, 158
 - value of, 4–5
- DCAT, 42, 175, 179
- decision support systems, 47
- deep learning, 176, 178
- de-identification, 13
- Deutsche Bundesbank, 12, 15, 19
 - Integrated Microdata-based Information and Analysis System (IMIDIAS), 13
 - Research Data and Service Centre (RDSC), 12–19
- DICE (company), 8
- digital object identifiers (DOIs), 18–23, 30, 45
- digital objects, 41
- Digital Repository Audit Method
 - Based on Risk Assessment, 46
- 'direct representation' for research and scholarship, 49
- Dryad Digital Repository, 45
- Dublin Core, 42
- Earth Observation System, 44
- EarthCube portal, 44
- EDGAR thesaurus, 42
- entity linking, 172–3
- ethical considerations, 30, 32
- European Social Survey, 44
- European Union (EU), 42
- evidence-based policy, 40, 49, 168
- Evidence-Based Policymaking
 - Commission, US, 3
- exchanging data, 48
- Extended Knowledge
 - Organization System (XKOS), 42
- Facebook, 28
- FAIR data principles (findable, accessible, interoperable and reusable), 40, 170
- feedback loops, 15–16, 169–71
- Figshare, 33
- 'Five Safes' framework, 13
- Foundations of Evidence-Based
 - Policymaking Act (US), 3
- Friend of a Friend (FOAF)
 - ontology, 42–3
- funding constraints, 32–3
- FUSE programme, 43
- Generic Statistical Information Model (GSIM), 48
- GESIS (company), 8, 108
- Ghavimi, B., 18, 148, 158
- gold standard corpus (GSC), 6, 57
- Google Dataset Search, 45, 54
- governance of data, 175
- graph embedding, 176
- Gupta, S., 130
- Health Level Seven International, 45
- Hearst, M.A., 118
- heuristics, 131
- Hochreiter, S., 161
- identification of datasets, 18–21, 24, 54
- implicit entity linking, 144, 148
- INEXDA, 18, 25
- InfoLink system, 18
- information infrastructure, 5, 28–31
- information retrieval from
 - scientific articles, 130

- infrastructure *see* cultural infrastructure;
information infrastructure
- instruments, sensitivity of, 28
- Internet of Things, 49
- Inter-university Consortium for
Political and Social Research (ICPSR), 41,
44, 58–62, 73
- DDI-Lifecycle model, 46
- Jansen, T., 130
- journals, 24, 30
- KAIST (company), 7, 94
- keyphrases, 144
- knowledge graphs, 170–1, 174–9
 - iterative improvement of, 174
- knowledge-generating processes, 14–17, 24
- Kohlmeier, Sebastian, 7, 78
- Kuhn, T., 130
- Kyoto Encyclopaedia of Genes
and Genomes, 45
- Leibniz Institute for the Social Sciences, 108
- link prediction, 172
- linked data, 42, 168–9, 178
- Lu, M., 158
- Luan, Y., 144
- machine learning models, 172–3, 176, 178
- machine reading for question
answering (MRQA), 95–9
- Mahata, D., 130–1
- Manning, C., 130
- Marine Metadata Interoperability Ontology
Registry and Repository, 47
- Mathink, N., 18
- Mendel, Gregor, 2
- metadata, 4–6, 21, 25, 29, 33,
40–3, 46–50, 145, 168–70, 174–9
 - types of, 6
- metadata schemas, 18
- methods of research, 19, 22–3
- microdata, 12–14, 19, 21, 46, 168
 - linking, 17
 - user-centric, 25
- ‘named datasets’, 19–21, 24
- named entity recognition (NER) modules,
84–91, 98–100, 103, 110, 113–18
- namespaces, 42
- Nasar, Z., 114, 144
- Nathan, Paco (co-editor), 8
- natural language processing (NLP),
4, 55, 154, 160
- natural sciences, 44
- New York University, 158
- nodes, 22
- ontologies, 42–3, 46, 175; *see also* social
ontology
- open access, 32, 35, 40
- Open Archival Information System (OAIS),
45–6
- Open Biomedical Ontology (OBO), 43
- open-source software and open standards,
55, 168–71, 174–6, 179
- Organization for Economic Cooperation and
Development (OECD), 42, 44
- peer review, 29
- personal data, 28
- Philosophical Transactions* (journal), 29
- positive externalities, 16
- predictions, 23
- privacy compliance, 46, 169, 178
- prizes for innovation, 54–5
- Project Jupyter, 176, 179
- Protein Data Bank, 44
- publication citation networks, 23
- ‘pull incentives’, 54
- ‘push programmes’, 54
- randomized controlled trials (RCTs), 20
- ReadCube, 33
- recommendation of data sources, 17, 20–4,
54, 178
- recurrent neural networks (RNNs), 161
- Registry of Research Data Repositories, 45
- reinforcement learning, 178
- repositories, 6, 40, 43–9
- reproducibility of research, 24, 29, 168
- research data centres, 24
- research fields, 19, 22
- researchers
 - collaboration between, 169
 - focused on data science, 35
 - motivation of, 31
 - retention of, 31
- rich context, 169–78
 - case in favour of, 169–71
 - core tenets of, 170
- in datasets, 3–5
 - definition of, 54
 - goal of work with, 170
 - use cases for, 171–6
- ‘Rich Context’ Competition (RCC)
(2018–2019), 4–9, 12–13, 28, 33, 36–7,
54, 76–8, 84, 90, 108, 111, 114–15, 124,
144–5, 153, 158, 162, 171–9
- building and submission of models, 65–9
- design of, 56–65
- evaluation of entries, 69–76, 80–1
- lessons learned from, 18–24
- objective of, 56, 108
- ongoing project to follow, 171

- outcome of, 171
- scientific challenges addressed in, 4
- timelines for, 9
- richer semantics, 42–3
- Ritchie, Felix, 13
- Royal Society, 29
- Rule-Oriented Data System, 46
- safe projects, safe people, safe data, safe settings and safe output, 13
- Sage Encyclopaedia of Social Research Methods*, 117
- Sage Publications, 58–9, 117–19, 139, 144–5, 148–50
- Schema.org, 41–2, 45
- Schmidhuber, J., 161
- Schwartz, A.S., 118
- Semantic Scholar*, 7, 79, 84
- semantic models, 48–9
- sensitive data, 169, 175, 178
- sharing of data, 32, 37, 169
- Simple Knowledge Organization System (SKOS), 42
- Singapore Management University, 8
- Singhal, A., 158
- Soboroff, Ian, 55
- social ontology, 43
- Social Science Open Access Repository (SSOAR), 18, 110, 118–21
- social sciences, 3, 28, 32, 168–72
- societal benefits, 24
- societal knowledge, 17–18
- Søgaard, A., 144
- Srivastava, J., 158
- statistical data collection, 44
- Statistical Data and Metadata eXchange (SDMX) guidelines, 40
- statistical methods, 23
- statistical models, complexity of, 168
- stewardship of data, 169, 174, 179
- supervised or semi-supervised learning*, 79, 173
- support vector machine (SVM) classifiers, 148
- Syracuse University, 8
- Taverna workflow tool, 48
- text analysis tools, 24–5
- text-mining, 20
- theoretical research, 22
- thesauri, 22–3
- tiered access to data, 169
- TopQuadrant, 175
- transitive inference, 173
- Triplestores, 49
- trust in data, 30
- UniProt, 44
- United Nations, 44
- United States
 - National Aeronautics and Space Administration (NASA), 44
 - Office of Science Technology and Policy, 44
- user engagement, 5
- versioning of data, 29, 33
- virtual research environments, 47
- visualization technologies, 29
- VIVO, 42
- vocabulary, 133
- Wade, Alex, 7, 78
- warehousing of data, 48
- Web Ontology Language (OWL), 43
- Westergaard, D., 130–1
- Wikidata, 42
- WissKI platform, 47
- workflows, 48, 169, 175, 178
- World Bank, 42, 44
- World Health Organization (WHO), 112
- W3C Data Activity, 175
- Zhang, Q., 18

