Rich Context

The social sciences are at a crossroads. The enormous growth of the scientific enterprise, coupled with rapid technological progress, has created opportunities to conduct research at a scale that would have been almost unimaginable a generation or two ago. The rise of cheap computing, connected mobile devices, and social networks with global reach allows researchers to rapidly acquire massive, rich datasets; to routinely fit statistical models that would once have seemed intractably complex; and to probe the way that people think, feel, behave, and interact with one another in ever more naturalistic, fine-grained ways. Yet much of the core infrastructure is manual and ad-hoc in nature, threatening the legitimacy and utility of social science research.

We can and must do better. 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 social science to understand the sources and consequences. Yet the lack of reproducibility and replicability evident in many fields(*1–5*) is even more acute in the study of human behavior both because of the difficulty of sharing confidential data and because of the lack of scientific infrastructure. The central argument we advance in this monograph is that advances in technology—and particularly, in automation—can now change the way in which social science is done. Social scientists have eagerly adopted new technologies in virtually every area of social science research—from literature searches to data storage to statistical analysis to dissemination of results.

A major challenge is search and discovery. The vast majority of social science data and outputs cannot be easily discovered by other researchers even when nominally deposited in the public domain. A new generation of automated search tools could help researchers discover how data are being used, in what research fields, with what methods, with what code and with what findings. And automation can be used to reward researchers who validate the results and contribute additional information about use, fields, methods, code, and findings.(*6*)

In sum, the use of data depends critically on knowing how it has been produced and used before: the required elements what do the data ***measure***, what ***research*** has been done by what ***researchers,*** with what ***code***, and with what ***results***.
Acquiring that knowledge has historically been manual and inadequate.

The challenge is particularly acute in the case of confidential data on human subjects, since it is impossible to provide fully open access to the source files.

This monograph provides pathbreaking contributions of different approaches to automating the collection and codification of knowledge from publications and people. Each paper summarizes technological approaches to applying text analysis techniques on a series of different publication corpora to identify the datasets referenced in each publication and draw out the required elements. The authors have been identified as a result of an international competition that challenged computer scientists to find ways of automating the discovery of research datasets, fields & methods behind social science research publications.. They are 1. [GESIS]{.s1}: Wolfgang Otto, Katarina Boland, Dimitar Dimitrov, Behnam Ghavimi, Narges Tavakolpoursaleh, Andrea Zielinski, Karam Abdulahhad 2. [KAIST]{.s1}: Haritz Puerto San Roman, Hong Giwon, Cao Minh Son 3. [Paderborn University]{.s1}: Rricha Jalota, Nikit Srivastava, Daniel Vollmers, René Speck, Michael Röder,Dr. Ricardo Usbeck, Prof. Dr. Axel-Cyrille, Ngonga Ngomo 4. [Allen AI]{.s1}: Waleed Ammar, Christine Betts, Daniel King, Iz Beltagy

We also will have a contribution from Daniel Acuna, Syracuse University

We envision additional contributions from the technical judges as well as the social science judges.

1. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637 (2018). 2. A. Dafoe, Science deserves better: the imperative to share complete replication files. *PS Polit. Sci. Polit.* **47**, 60–66 (2014). 3. J. P. A. Ioannidis, Why Most Published Research Findings Are False. *PLoS Med*. **2**, e124 (2005). 4. N. Young, J. Ioannidis, O. Al-Ubaydli, Why Current Publication Practices May Distort Science. *PLoS Med* (2008). 5. G. Christensen, E. Miguel, Transparency, reproducibility, and the credibility of economics

research. *J. Econ. Lit.* **56**, 920–980 (2018). 6. T. Yarkoni *et al.*, "Enhancing and accelerating social science via automation: Challenges and Opportunities" (2019).

Placeholder between chapters

Chapter Break

Placeholder between chapters

Chapter Break

Title: Who’s Waldo: Conceptual issues when characterizing data in empirical research Author:Stefan Bender, Hendrik Doll, Christian Hirsch [1] Affiliation: Research Data and Service Centre , Deutsche Bundesbank Date: June 14, 2019

Abstract

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

Table of contents

Introduction

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

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

Empirical research papers are an obvious source of information about dataset usage. A useful microdata recommendation system needs to rely on a corpus of dataset usage, as large as possible. Hand-curating such a sufficiently large corpus is prohibitively labor-intensive and error-prone. Therefore, being able to automatically retrieve the necessary information from research papers lays the groundwork for any future implantation of such a recommendation system. The competition is an important first step and proofs data set extraction from research publications to be feasible and scalable.

We present lessons learned from the machine learning competition held to automate the discovery of datasets, associated research methods and fields in social science research publications. In doing so, we show our insights about dataset taxonomies from our experience in a research data center and from designing a machine learning competition. We do this with a background of all authors in social science. We refer the readers interested in the more technical aspect of the task of extracting dataset citations from publications to later chapters.

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

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

Insights from a research data centre perspective

Background

The overriding principle of Bundesbank – and other Central Banks, National Statistical Institutes and Of-ficial International Institutions - when working with micro data is compliance with the statutory secrecy and data protection requirements, and thus maintaining the confidentiality of the information submitted by the reporting agents. European and national legal provisions regulate both the user group and the access channels to micro data, prescribe the required degree of data anonymisation and oblige data providers and data recipients to maintain data confidentiality at all times.

In response to the increased internal and external demand for microdata and the data confidentiality re-quirements, in 2013 the Bundesbank set up the Integrated Microdata - based Information and Analysis System (IMIDIAS) and established the Research Data and Service Centre (RDSC) (for a detailed moti-vation, refer to Kalckreuth, 2014 and Bender and Staab, 2015). The RDSC applies a standardised pro-cedure to generate high-quality datasets that cover a large part of data requests for research purposes. Thereby, the RDSC grants internal and external researchers' access to selected Bundesbank microdata and serves as an interface between data producers and data users.

Requests to use microdata are first reviewed pursuant to legal requirements. The RDSC provides access to anonymized datasets on banks, securities, investment funds, enterprises and households, all of which can be accessed at dedicated researcher workstations or for most of the Bundesbank's surveys – as for the Panel on Household Finances (PHF) study – the RDSC offers so called scientific use files. Data access and the underlying legal requirements are described in detail by Schönberg (2018).

In addition, the RDSC provides information and advice to researchers on data selection, data content and analytical approaches. Together with the relevant statistical experts, it ensures that the microdata provided are documented in detail and archived. In doing so, the RDSC works according to globally rec-ognized standards and was accredited as a research data centre (RDC) by the German Data Forum ("Rat für Sozial- und Wirtschaftsdaten").

To date, metadata in the RDSC is provided to research using structured data reports. They are an es-tablished and well-functioning tool to convey information linearly from the data producers via the RDSC to the data users. To go a step further, efforts are underway in the RDSC to document microdata from the usage side. This relevant source of information has not been considered yet, because necessary in-formation on which datasets are used in which publication has not been broadly available to date. The competition changes this for the first time.

Potential of such data documentation from the usage side is manifold. Examples include newly arriving researchers, who get to see, what other researchers did with the data. At a glance, one can directly see potential for linkages with other micro datasets if others have done it. Data producers benefit from feed-backs on potential data gaps or limitations of the used data (which goes back to the initial institutional motivation for creating a research data center). Thereby a circular information flow is created by allowing feedback loops.

A main principle of the RDSC is to give free access to Bundesbank micro data for independent research. Motives for doing so are to get feedback on the data (use published research results to increase the internal data knowledge) and to strengthen evidence-based policy-making by Bundesbank itself. For fulfilling both of these tasks, the RDSC has to ensure microdata is used effectively by providing excellent services. Implementing potential for structured feedback from researchers back to data production and new research enables an improving empirical knowledge generating process.

The knowledge generating process of empirical research in the RDSC

The knowledge generating process of empirical research in the RDSC can be organised along the four key dimensions (i) data services, (ii) research, (iii) publication, and (iv) (structured) user specific knowledge.

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

Data Services comprises raw microdata and comprehensive documentation of the data both of which the RDSC compiles together with the data producing units in Bundesbank. [2] Furthermore, the data ser-vices dimension also includes the methodological improvement of microdata through e.g. applying rec-ord linkage techniques to facilitate the creation of new datasets for research. Finally, the RDSC also of-fers advisory services to potential and existing microdata users on topics such as e.g. dataset selection or analytical options.

The second dimension of the knowledge generating process of empirical research in the RDSC is re-search. After the application of a researcher is approved by the RDSC, researchers conduct their re-search 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 anonymised research outcomes may be used outside of the secure environment.

Researchers as users of data services produce research outcomes. These outcomes – after data confi-dentiality clearance – sometimes take the form of publications, which present results to the interested public in a form optimized for human consumption as unstructured text. These publications contain knowledge accumulated by researchers about data usage over time (experience), e.g., knowledge about dataset particularities, which in turn could be utilised 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, us-ing additional information) • What purposes data is used for (e.g. topics, methodology, research area) • What kinds of analyses or techniques have been tried and are used ultimately • What information about data is most valuable to get to the results, respectively which linkage or data enrichment makes renders the data most valuable.

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

The challenge is to establish such a feedback loop. If effective feedback is given and used, the microda-ta-based knowledge-generating process restarts with data services, but on a higher level. Better data services in turn allow better research, because available microdata is better described and more effec-tively used. By automatizing the feedback-loop between research, publications, knowledge, and data services, the knowledge generating process can loop faster and augment quicker. We expect this to lead to improvements in the four key dimensions of this process of empirical research in the RDSC.

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

At the moment, this feedback loop is not present in a systematic way. The aim of the competition is to identify appropriate procedures to close the gap between publication and data services, which would enable transforming knowledge available in publications into generally re-usable knowledge to inform stakeholders (data producers, RDSC, decision makers at Bundesbank). The results of the competition will thus ultimately enable better data services which in turn will make research outcomes more efficient through the channel of a more optimal 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 recommen-dations and data documentation). Second, the case for societal investment in free data access can be empirically fortified. Positive externalities (i.e. research as a public good) suggests a less then societally optimal provision of research data and related services. Obtaining a dataset impact factor can then make the case for further investment in microdata provision by concretely showing a dataset's impact.

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

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

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

In our setting, with the knowledge generating process of empirical research in the RDSC, we consider researchers and datasets. The universe of data users/ researchers is decently large (i.e. the first dimen-sion), but per user, we only observe a limited amount of "dataset consumption" (i.e. the second dimen-sion). Hence, we have a decent chance of recommending based on other users behaviour. However, we have only limited means of predicting a single users 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, less data points per person are needed, in order to make sensible recommendations. Also, in order to gain more of the rare information per user, we can fall back on dataset citations, i.e. "indirect data usage", as outlined in chapter 3.

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

Second, the RDSC as part of a public institution has a responsibility towards its principals i.e. society. Granting data access free of charge for researchers should be backed by empirically measurable bene-fits of such data provision. Benefits from data usage can justify societal investment in free data access. However, measuring societal benefits through data access is not obvious at first glance. One possible starting point of approximating societal benefits of data access can be to measure the creation of knowledge [3] created by specific datasets.

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

Lessons learned from competition

Related literature

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

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

Zhang et al (2016) also use a bootstrapping approach to extract dataset citations from 116 computer science journals publications. Ghavimi et al. (2016) use a similar approach for social science papers finding datasets with well-documented metadata. According to them, only 25% of all dataset citations are given in the references, highlighting the unstructured citation culture for datasets. We advance from these with an environment with less available dataset metadata and a corpus of publications from a va-riety 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 DataCite metadata schema and the da|ra metadata schema, which complies with the DataCite schema (Helbig et al. 2014). They offer dataset taxonomies and standardized citation propositions, however their categories do not optimally support automatic search and extraction, if no unique dataset identification (such as a DOI) is used.

In the con-text of central banks that provide microdata, recent progress has been made in the context of INEXDA. A metadata standard (in line with DataCite) has been developed (Bender et al. 2018) and datasets pro-vided by the RDSC are all DOI registered.

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

Dataset mentions

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

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

Lesson #1: datasets fall into two broad categories

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

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

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

In contrast to named datasets, created datasets usually are not referred to by a specific string or com-monly used abbreviation. Data collection is usually paper specific, and the universe of existing datasets are not easily searchable. This makes it hard for text mining algorithms to correctly extract strings refer-ring to dataset entities. Specific created datasets are harder to use for follow-up research, and reproduc-ibility is given only if publishers provide data together with the paper. Therefore, the lack of unique iden-tification 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 duration 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 identifica-tion techniques and differing potentials for conducting RCTs, the predominantly used data sources natu-rally vary. This has important repercussions for designing a competition, since algorithm performance and later recommendation system performance varies with the input corpus and the application field.

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

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

Lesson #3: Unique identification of datasets remains an issue

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

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

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

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

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

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

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

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

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

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

Fields and Methods

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

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

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

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

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

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

Discussion

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

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

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

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

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

Conclusions

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

We introduce a circular model of the knowledge generating process, which increases in levels. From da-ta 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. Thereby the circle repeats on a higher level. The current competition works on strengthening the knowledge pillar as well as the transmission mechanisms from publications to knowledge to improved data services. [4]

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

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

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

References

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

Placeholder between chapters

Chapter Break

Placeholder for Dimensions use case chapter.

\pagebreak

Placeholder between chapters

Chapter Break

Rich Context Book Chapter - Standardized Metadata, Full Text and Training/Evaluation for Extraction Models

[**Standardized Metadata & Full Text [Sebastian]**]{.s1}

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

One way these challenges can be addressed is by using a trustworthy metadata source like [[Semantic Scholar's open corpus]{.s2}](http://api.semanticscholar.org/corpus/) developed by the Allen Institute for Artificial Intelligence (AI2) or [[Microsoft's Academic Graph]{.s2}](https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema) that make it easy to access standardized metadata from an openly accessible source. In addition, both Semantic Scholar and the Microsoft Academic Graph provide topics associated with papers which makes it easy to narrow down papers by domain. If full text is needed we recommend tying the metadata to a source of open access full text content like [[Unpaywall]{.s2}](https://unpaywall.org/data-format) to ensure that the full text can be freely redistributed and leveraged for model development.

To gather the data we recommend collecting a sufficiently large set of full text papers (3,000–5,000 minimum) with their associated metadata and providing participants with a standardized format of the full text. More data might be required if data is split across many scientific domains. For example for a task like dataset extraction, reference formatting is often inconsistent across domains and dataset mentions can potentially be found in different sections (e.g. background, methods, discussion, conclusion or the reference list) throughout the text. Once a decision has been made on the full text to include, the PDF content can be easily converted into text in a standardized format using a PDF to text parser like [[AI2's ScienceParse]{.s2}](https://github.com/allenai/spv2) (which handles key tasks like metadata, section heading and references extraction).

Once the metadata and full text dataset has been created it can be easily versioned and used again in future competitions. For example, if updated metadata is needed it's easy to go back to the original metadata source (for example by using Semantic Scholar's [[API]{.s2}](http://api.semanticscholar.org/)) to get the latest metadata.

[**Annotation Protocols to Produce Training & Evaluation Data [Alex]**]{.s1}

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

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

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

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

One way to evaluate how well your model performs is to measure the **recall** and **precision** of the model's output, and in order to do this we can compare the output to the labelled evaluation subset. In other words, how well does our model perform against the human annotations that it was not trained on and has never seen. Recall is the percentage of right answers the model returned. For example, if the evaluation dataset contained 1000 mentions of a dataset, and the trained model returned 800 of them, then the recall value would be .80.
But what if the model returned everything as a

dataset, then it would get all 1000, plus a whole bunch of wrong answers. Obviously, the precision of the model is important too.
Precision is the percentage of answers returned that were right. So, continuing the example above, if the model returned 888 answers, and 800 of those were right, then the precision of the model would be ~.90.
But again, if the model returned only one right
answer and no wrong ones, the precision would be perfect. So, it is important to measure both precision and recall.
In summary, the model in this example, got 80%
of the right answers, and 90% of the answers it returned were right. The two measures of recall and precision can be combined into an F1 score of ~.847.\
If we then make modifications to our model, we can re-run it against the evaluation dataset and see how our F1 score changes. If the score goes up, then our new model performed better against this evaluation data. If we want to compare several different models to see which one performed best, we can calculate an F1 score for each of them. The one with the highest F1 score has performed the best. Consequently, the quality of the annotations are critical for two reasons: first, the accuracy of a *model* will only be as good as the data upon which it was trained. And secondly, the accuracy of the *evaluation* (in this case the F1 score) can be affected by the quality of the data it is evaluated against.

\

Placeholder between chapters

Chapter Break

Metadata for Administrative and Social Science Data

Robert B Allen

[0000–0002–4059–2587]

rba\@boballen.info

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

INTRODUCTION

Evidence-based policy needs relevant data (Commission on Evidence-Based Policy, 2018; Lane, 2016). Such data is valuable, but often difficult to find and/or replicate. Therefore, data stewardship is needed. The FAIR Open Access guidelines suggest that, ideally, data should be Findable, Accessible, Interoperable, and Reusable.[5]

Data may be a text, an image, or a video but this chapter focuses on numeric observations recorded and maintained in machine-readable form. There are many data sets available online; the DataCite[6] repository alone contains over five million. There are many different types of data sets. Data sets differ in their structure, their source, and their use. 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 data sets. Following the approach of W3C-DCAT (World Wide Web Consortium-Data Catalog Vocabulary)[7], a data set may be a collection of related observations which is developed and managed by a single entity such as a statistical agency. When stored as a unit online, the data set is a digital object.

Metadata consists of short descriptors which refer to a digital object. Metadata can support users in finding data sets, and enable users to know what is them. 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. Structural metadata includes the organization of the files. Administrative metadata describes the permissions, rights, and usage. Descriptive metadata covers the contents.

This chapter surveys the state of the art of metadata for data sets, focusing on metadata for administrative and social science records. Administrative records describe details about the state of the world. They could include governmental records, hospital records, educational records, or business records. By comparison, social science data often involves a theory. In some cases, the theory is applied to interpreting the data or the data are used to develop theory.

Section 2 describes data, metadata, and digital objects. Section 3 discusses semantics. Section 4 considers repositories. Section 5 describes services. Section 6 describes the techniques for documenting the internal structure of data sets. Section 7 discusses cyberinfrastructure. Section 8 closes with recommendations and conclusions.

DATA, METADATA, AND DIGITAL OBJECTS

A metadata element describes some attribute of a digital object. The simplest metadata identifies the digital object.[8] Individual metadata elements are generally part of a set which describes attributes of a data set. Such a set of metadata elements can be structured as a catalog, schema, or frame, and restrictions can be placed on the values allowed for the individual elements. A fragment of an example of the Schema.org[9] dataset schema is shown in Figure 1. Note the distinct metadata elements in that fragment.

Figure 1: Fragment of schema.org dataset schema[10].

The W3C DCAT[11] is a schema for data sets that is used by many repositories such as data.gov. There are still other descriptive frameworks for data sets such as the DataCite[12] metadata schema and the Inter-university Consortium for Political and Social Research Data Documentation Initiative (ICPSR DDI) discussed below (Section 4.1). The catalog specifications provide a flexible framework. For instance, DCAT allows the inclusion of metadata elements drawn from domain schema and ontologies. Some of these domain schemas are widely used resources which DCAT refers to as assets. For instance, spatial relationships are often modeled by the Federal Geographic Data Committee (FGDC) standard.[13]

Many of the implementations for indexing collections of metadata schemas use relational databases. Thus, they use SQL and support tools such as data dictionaries. Moreover, they are often characterized by Unified Modeling Language (UML) Class Diagrams which are common for data modeling. Other implementations use triplestores which have comparable tools.

SEMANTIC DESCRIPTIONS

Semantic data models have become widely explored. In the Semantic Web, nodes are implemented with XML. RDF (Resource Description Framework) asserts a relationship between two identifiers by defining a group of three nodes (triples) as “identifier”-“property”-“identifier”. By connecting triples, RDF can define or graph a network of the relationships among a set of controlled vocabulary terms. This is the essence of linked data. RDFS (RDF Schema) extends RDF by supporting class/subclass relationships. The types of classes for identifiers in triples are controlled by domain and range parameters. Basic thesauri have a simple hierarchical structure. The Simple Knowledge Organization System (SKOS) is an RDFS standard for representing thesauri. Many administrative and social-science-related thesauri such as EDGAR, the World Bank, and the OECD have now been implemented with SKOS. A knowledge graph, a model of a domain, sometimes including instances, which is implemented in SKOS. Thus, DBpedia[14] is a knowledge graph based on Wikipedia.

Some frameworks for structural descriptions of data sets may include aspects of ontologies. Less formal ontologies simply provide definitions and employ RDFS. For example, Schema.org schemas can be used with micro-formats which match schema elements with passages in an online text. Schema.org has a classification of topics and may incorporate other systems such as FOAF (Friend of a Friend) which includes attributes associated with people. Formal ontologies use OWL (Web Ontology Language) to add features to RDFS. These features lend themselves to logical inference provided that the entities and relationships are rigorously defined.

Upper ontologies provide top-down structures for the types of entities allowed in derivative domain and application ontologies. One of the best known upper ontologies is the Basic Formal Ontology (BFO) (Arp, Smith, & Spear, 2015), which is a realist, Aristotelian approach. At the top-level, BFO distinguishes between Continuants (endurants) and Occurrents (perdurants) and also between Universals and Particulars (instances). Many biological ontologies have been developed based on BFO and are collected in the Open Biomedical Ontology (OBO) Foundry.

There are fewer rich ontologies dealing with social science content than there are for natural science. One challenge is “social ontology” that is in developing definitions for social terms. It is difficult to define exactly what is a family, a crime, or money. In most cases, an operational definition or an approximate definition may suffice where structured documentation of the definitions are unavailable. Moreover, while social terms are especially difficult to define for casual speech, it seems possible to make clear, though perhaps cumbersome, definitions for scholarly applications.

DATA REPOSITORIES

A data repository holds data sets and related digital objects. It provides access to the data sets and supports search. Metadata is integral to these services at several levels. In addition to item-level metadata for the data sets, there can also be study-level metadata or collection-level metadata.

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

ICPSR[15] is a major repository of public-use social science and administrative data sets derived from questionnaires and surveys. The ICPSR DDI[16] (e.g., Vardigan, Heus, & Thomas, 2009) defines a catalog code. A notable feature is a codebook which saves the exact wording of all the questions. In addition, the ICPSR provides an index of all variable names that are used in the data sets. DDI-Lifecycle is an extension of DDI that describes the broader context in which the survey was administered as well as the details about the preservation of the file (see Section 5.3).

Repositories of Governmental and NGO Statistical Agencies

Statistical data collection is a core function of government. Most countries have national statistical agencies. While these statistical collections often emphasize social data, they also include related indicators such as agricultural and industrial output and housing, such as Statistics New Zealand, and the Korean Social Science Data Archive (KOSSDA). European data sets are maintained in the Consortium of European Social Science Data Archives (CESSDA)[17] and the European Social Survey.[18] Australia has a broad data management initiative, ANDS.[19] Many U.S. governmental data sets are collected at data.gov.[20] In addition, there are many non-governmental and inter-governmental agencies such as the OECD, the World Bank, and the United Nations, which host data sets.

Other Data Repositories

Many data sets are produced, curated, and used in the natural sciences such as astronomy and geosciences. Some of these data sets have highly automated data collection, elaborate archives and established curation methods. Many of these repositories include multiple data sets for which access is supported with portals or data cubes (see Section 6.1).

For instance, massive amounts of geophysical data and related text documents are collected in the EarthCube[21] portal. The Science.gov portal is established by the U.S. Office of Science Technology and Policy. NASA supports approximately 25 different data portals. Each satellite in the Earth Observation System (EOS) may provide hundreds of streams of data,[22] with much common metadata. This provides a context analogous to study-level metadata.

Likewise, there are massive genomics and proteomics data sets which are accessible via portals such as UniProt[23] and the Protein Data Bank[24] along with suites of tools for exploring them. Similarly, there are very large data sets from medical research such as from clinical trials and from clinical practice including Electronic Health Records (EHRs).

Ecosystem of Texts and Data Sets

Data sets are often associated with text reports. For example, the Dryad Digital Repository[25] hosts data sets from scholarly publications which require that when a scholarly paper is accepted for publication that data associated with the paper is deposited.

Data sets may be cited in much the same way that research reports are cited. Formal citation facilitates tracing the origins of data used in analyses and helps to acknowledge the work of the creators of the data sets. Standards have been developed for such citations (Martone, 2014; Silvello, 2017).

SERVICES

The purpose of metadata and other aspects of information management is to provide services to users. Indeed, “service science” is an approach in information technology which focuses on the design and delivery of services rather than on underlying technologies.

Search

Searching for data sets differs from the familiar web-based text search because data repositories are generally hosted by either relational databases or semantic triplestores. Even where the data are stored on separate servers the metadata can be harvested and searched. This type of federated search is supported by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH);[26] both data.gov and ICPSR use OAI-PMH.

From Statistical Packages to Observatories

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, RSpace, and eLab notebooks (ELN) integrate annotations, workflows, raw data, and data analysis into one platform. In addition, some repositories have developed their own powerful data exploration tools such as ICPSR Colectica[27] for DDI and the GSS Data Explorer[28].

Virtual research environments (VREs) are organized by research communities to coordinate data sets with search and analytic tools. For instance, the Virtual Astronomy Observation (VAO) uses Jupyter to provide users with a robust research environment. WissKI[29] is a platform for coordinating digital humanities data sets which is based on Drupal.

Preservation

Lost data is often irreplaceable. Even if the data is not entirely lost, users need confidence that the quality of stored data has not been compromised. Moreover, although data storage prices are declining dramatically, we cannot save everything and the cost of maintaining a trusted repository remains substantial. Many of these challenges are familiar from traditional archives. For instance, selection policies could help in controlling the many poorly documented data sets in some repositories. Yet, prioritization is difficult[30] (Whyte & Wilson, 2010).

The Open Archival Information System (OAIS) provides a reference model for the management of archives (Lee, 2010). The OAIS framework has been incorporated into the ICPSR DDI-Lifecycle model. The Integrated Rule-Oriented Data System (iRODS)[31] is a policy-based archival management system developed for large data stores. It implements a service-oriented architecture (SOA) to support best practices established by archivists.

Provenance records the history of an entity. This can help to ensure confidence in its authenticity. For data in a repository, provenance often means tracing the history of repository operations. The history of transitions is often recorded as event data, where the events are what happened to the data in the dataset. Typically, provenance ontologies include actors, events, and digital objects. Potentially, blockchains could provide an even greater level of trust in digital provenance.

Metadata Quality and Metadata Management

Metadata, whether for texts or data sets, needs to be complete, consistent, standardized, machine processable, and timely (Park, 2009). A metadata editor supports the assignment of quality metadata (e.g., Gonclaves, O'Conner, et al., 2019). When collections or metadata standards change, the repository librarian must revise metadata (Tonkin, 2009). This might be particularly needed when updating metadata from data streams[32] such as those from satellite downlinks or from smart-city sensors.

Some repositories of survey data include micro-data, data for the responses that individuals gave to survey questions.[33] Although survey results are generally aggregated across individuals, individual-level data can sometimes very useful. Currently, there are no distinct metadata tags for such data; they are embedded into repository data. Moreover, the individual level of analysis raises privacy concerns and needs to be carefully managed. At the least, access should be limited to qualified researchers.

Metadata registries, such as the Marine Metadata Interoperability Ontology Registry and Repository,[34] record usage. The Registry of Research Data Repositories (re3data registry),[35] which is operated by DataCite, links to more than 2000 different repositories each of which holds many data sets. Each of the repositories is described by the re3data.org schema for the description of research data repositories (Rücknagel, Vierkant, et al., 2015).

Metadata application profiles provide constraints on the types of entities that can be included in the metadata for a given application. For instance, DCAT Application Profiles (DCAT-AP) support standardized metadata exchange between repositories in different jurisdictions in the EU.[36]

DETAILS ABOUT THE DATA IN DATA SETS

Data Cubes

Many notable data management techniques were originally developed for managing and processing business data.[37] Data cubes support access to multidimensional data. They present data as if it filled cells of a high-dimensional cube, even if the data will probably not fill all of the cells. Users can generate different views of the data by drilling-down, rolling-up, and slicing-and-dicing across cells. For complex data sets, there will be many dimensions. 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.

Data cubes have been extended beyond business information processing to cubes such as the Statistical Data and Metadata eXchange (*SDMX) used in the financial services industry and the* W3C Data Cube[38] standard that is applied in projects such as EarthCube.

Sequential Activities and Modeling Research

Entities change over time, yet knowledge representation frameworks rarely model change. In order to represent these changes, models need to represent transitions, processes, and other sequential activities. However, modeling change is routine for state machines, Petri nets, and other transition models. Such modeling is closer to the Unified Modeling Language (UML) or even programming languages than to ontologies. A "model-layer" that allows general statements to be made about sequential activities could incorporate both ontology and transitionals (Allen & Kim, 2018).

Models of sequential activities include workflows and mechanisms (e.g., Allen, 2018). A workflow is a structure for managing a sequence of activities and is a natural fit for describing research methods and analyses (Austin, Bloom, et al., 2017). The Taverna workflow tool has been widely used in the MyExperiment[39] project, which provides a framework for capturing and posting research methods and incorporates simple ontologies such as FOAF.

Allen (2015, 2018) has proposed direct representation of entire research reports. This approach uses a programming language that blends upper ontologies with object-oriented programming to do semantic modeling. Potentially, social mechanisms (e.g., Hedstrom & Ylikoski, 2010) and community models could be implemented. In addition to modeling events, it is also possible to use structured argumentation and assertions made in scientific research reports. Further, highly-structured evidence and claims might be applied to the evaluation of evidence-based social policy.

CYBERINFRASTRUCTURE

Information Institutions and Organizations

Libraries and archives (whether traditional or digital) have the mission, and often the resources, to develop standards and maintain information over the long-term. As noted earlier (Section 5.3), preservation is the fundamental concern for archival collections. Information institutions often have formal collection management strategies, metrics, and policies. These include Web and repository metrics and analytics, usage statistics such as reports of how many downloads were made from data sets, and procedures for updates and formatting standards.

In addition to traditional information institutions, there are now many additional players. These new players have slightly different mandates. For example, Schema.org's primary mission is to provide a structure that can improve indexing by search engine companies. Nonetheless, they often adopt the best practices similar to those of more traditional information organizations.

CrossRef[40] 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. Increasingly, the two projects are coordinating. ORCID iDs[41] are persistent digital identifiers assigned to authors. The emergence of structured identifiers such as DOIs and ORCID iDs has allowed the development of services such as VIVO[42] and the Microsoft Academic Graph (MAG)[43] which allow authors to be tracked across research reports and projects, and across publishers.

Cloud Technologies, the Internet of Things, and Software as a Service

We are now well into the era of cloud computing (Foster & Gannon, 2017), allowing flexible allocation of computing, networking and storage resources. Networked data centers needed for cloud computing facilitate the Internet of Things (IoT). Cloud computing also facilitates Software as a Service (SaaS).

The compatibility of the versions of software packages needed for data management is often a challenge. Containers, such as those from Docker, allow compatible versions of software to be assembled and run on a virtual computer. A container could hold datasets, and workflows, as well as the programs used to analyze the data, making the analyses readily replicable.

RECOMMENDATIONS AND CONCLUSIONS

Some of the biggest issues for the retrieval of information from data sets concern information organization. Metadata supports the discovery of and access to data sets. While there is already a lot of attention to metadata, even more attention would further support evidence-based policy. We need richer, more systematic, and more interoperable metadata standards.

ACKNOWLEDGMENTS {#acknowledgments .ListParagraph}

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

REFERENCES {#references .ListParagraph}

Allen, R.B. (2015). Repositories with direct representation, *Networked Knowledge Representation Systems,* arXiv: 1512.09070

Allen, R.B. (2018). *Issues for Using Semantic Modeling to Represent Mechanisms*, arXiv:1812.11431

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

Arp, R., Smith, B., & Spear, A.D. (2015). *Building Ontologies with Basic Formal Ontology*, MIT Press, Cambridge. MA.

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

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

Foster, I., & Gannon, D.B. (2017). Cloud Computing for Science and Engineering, MIT Press, Cambridge, MA.

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

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36* 49–67.

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

Lee, C.A. (2010). Open Archival Information System (OAIS) Reference Model. *Encyclopedia of Library and Information Sciences* (3rd Edition). CRC Press, doi: 10.1081/E-ELIS3–120044377

Martone M. (ed.) (2014). *Joint Declaration of Data Citation Principles*. San Diego CA: FORCE11, Data Citation Synthesis Group: https://www.force11.org/group/joint-declaration-data-citation-principles-final

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

Rücknagel, J., Vierkant, P., Ulrich, R., Kloska, G., Schnepf, E., Fichtmüller, D. et al. (2015): Metadata Schema for the Description of Research Data Repositories: version 3.0 (29), doi: 10.2312/re3.008

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology, 69*(1) 6–20. doi: 10.1002/asi.23917

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs**, R.R.,** Duerr, R., et al. (2015). Achieving human and machine accessibility of cited data in scholarly publications, *PeerJ Computer Science* 1: e1, doi 10.7717/peerj-cs.1

Tonkin, E., (2009). MetRe supporting the metadata revision process.
International Conference on Digital Libraries,

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

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

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

Chapter Break

Placeholder for Bob Allen chapter on broader information context.

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

Jonathan Morgan designed and implemented the data annotation and evaluation strategies and the first draft of our model packaging framework. Andrew Gordon sampled and prepared the corpus for processing and for distribution to participants. Ekaterina Levitskaya helped with the design of refinement of coding protocols and did a substantial amount of the data annotation. Jonathan and Andrew worked together to collect and run submissions for the competiton and summarize the output for the judges.

Table of Contents

Introduction

The rich context competition was designed as the start to a series intended to inspire data scientists to use artificial intelligence and machine learning to develop and identify text analysis and machine learning techniques to discover relationships between data sets, researchers, publications, research methods, and fields.

The competition had multiple goals in addition to inspiring work on models to detect mentions of data in academic publications.

One goal was to use the competition and the workshop that concluded it to create a community of practice around the creation of rich context, and more specifically, around the creation of data, methods, and models needed to better detect data in publications, and so tie data to concrete outcomes.

Another goal was to inspire innovation in a space where technology had not been previously focused, and where we found data and infrastructure to support our task were lacking.

To meet these goals, we designed a contest that used prize money to inspire innovation, included group meetings and workshops to try to build a community of practice among participants, and made the resulting models and as much of the data for the competition as possible open source and freely available, so that anyone could build on the results. In addition, we are writing this book to provide participants with a publication and document the competition to make its design more readily reusable.

Previous Work

Incentivizing innovation

At a high level, most systematic incentives for innovation can be classified as one of two types: up-front support for research ("push programs") or commitments to reward successful results ("pull incentives") (Kremer & 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. Heidi Williams, in (Williams, 2012), provides an overview of the trade-offs inherent in innovation prizes: 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). The United States federal government Office of Science and Technology Policy (OSTP) Innovation Toolkit is an example of an institution trying to agree on appropriate parameters for innovation, including innovation prizes (Kalil & Miller, 2015).

The rich context competition aimed to not only inspire innovation, but also to build a community of practice that could grow and build on the work it incentivized. Communities of practice, particularly in a domain of work where cumulative knowledge is used to continually build on past advances, have many benefits: tendencies to develop knowledge-sharing and dissemination mechanisms, common norms of sharing and cooperation, and broad agreement on technical paradigms and jargon (Boudreau & Lakhani, 2009). As open source software communities show, however, they must be carefully incentivized and nurtured to grow participation (Mateos-Garcia & Steinmueller, 2008) and managed well to maintain resources and quality of output over time (Sadowski, Sadowski-Rasters, & Duysters, 2008).

To facilitate building on the output of the competition, we made a substantial subset of the data used for training and testing openly available, and required that all entries be documented well and be licensed under an open source license. This kind of openness lowers cost of accessing existing work, and allows groups to more readily identify and embark on novel research, rather than reproducing others' obscured work (Murray, Aghion, Dewatripont, Kolev, & Stern, 2016).

Competition Design

In Natural Language Processing (NLP), incentivizing innovation can be complicated. The challenges in deriving context in academic publications are similar to those outlined based on analysis of clinical text: access to shared data is limited, there are no existing annotated data sets or standards for annotations, existing solutions are not easily reusable, 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). Our competition aims to incentivize collective development of technologies for detecting the presence and use of data in publications similar to the FUSE project's work predicting success of technologies based on patents and papers in which they are discussed (Reardon, 2014).

In the NLP domain, Ian Soboroff at the National Institutes of Standards and Technology (NIST) has developed a series of competition patterns for inspiring disparate groups of researchers to help to carry out information tasks against text data. These include more 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, Ounis, Lin, & Macdonald, 2013) and more elaborate scenarios like one organized around an "incident", where groups are given training data and model specifications and allowed to train a model, then game out an incident where an event occurs in a previously unseen language and they have to quickly adapt their model to the new language and submit results (Tong et al., 2018).

Our competition was influenced by these designs, specifically building on the design of the 2015 PatentsView Inventor Disambiguation Technical Workshop (http://www.patentsview.org/community/workshop-2015), which was a machine learning competition in two phases with a cash prize of $25,000. In phase one, participants ran their models against a common set of open data provided to participants and were evaluated on performance against this open data. The top participants were then invited to phase 2, where their models were then run and evaluated against a holdout.

Data creation

For training and evaluation data, our goal was to lay the foundations for developing a “gold standard corpus” (GSC) of examples of language around data being mentioned and used in analysis in academic publications. A GSC corpus is one that is manually tagged and reviewed for quality, usually for a particular domain and task, and creating one is time-consuming and expensive (Wissler, Almashraee, Díaz, & Paschke, 2014). Wissler et al. outline options for decreasing the cost, including starting with a “Silver Standard Corpus” (SSC) created using chained machine learning models and annotation via crowd-sourcing, but in general you must select a corpus to annotate, then implement a manual annotation and review scheme.

While our goal was not to make a GSC, we used our data creation to begin to assess data needed for high quality data detection models and to test potential methods for creating a GSC. To create our competition training and evaluation data, we started with data set citation data from the ICPSR data catalog (https://www.icpsr.umich.edu/icpsrweb/), then used methods that originated in quantitative content analysis of communication artifacts (Riffe, Lacy, & Fico, 2005) combined with software designed to reduce and simplify the work of human coders to increase reliability (Lewis, Zamith, & Hermida, 2013).

Competition Design

The goal of this first round of 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 competition had two phases.

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

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

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

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

Data

In each of the two phases, competing teams were given text and metadata for 5,000 publications and single set of metadata on 10,348 data sets of interest, shared between the two phases, for use in training and testing their models. Separate 5,000-publication samples were provided for each phase. The corpus of 10,348 data sets included data maintained by

Deutsche Bundesbank and the set of public data sets hosted by the Inter-university Consortium for Political and Social Research (ICPSR). In addition, a single 100-publication development fold was provided separate from the training and testing data to serve as a test for packaging of each team's model, and as a quick test of their model and the quality of its output. For details on the metadata provided for each type of data, see [[https://github.com/Coleridge-Initiative/rich-context-competition/wiki/Dataset-Description]{.underline}](https://github.com/Coleridge-Initiative/rich-context-competition/wiki/Dataset-Description).

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

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

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

Publications

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

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

Publication Dataset - Phase 1

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

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

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

The 2,500 un-annotated publications used in the holdout evaluation corpus were sampled from across Sage Publications' journal holdings including non-open access journals.

Publication Dataset - Phase 2

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

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

Converting PDF files to plain text

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

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

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

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

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

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

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

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

Data Sets

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

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

- subjects - list of terms associated with the dataset, based on the
- [[ICPSR subject thesaurus.]{.underline}](https://www.icpsr.umich.edu/icpsrweb/ICPSR/thesaurus/subject)additional_keywords - System keyword for where dataset originated.
- citation - Preferred dataset citation.
- data_set_id - Integer ID for dataset from our internal data store of publications, data sets, and relations. This is the identifier used in the data_set_citations.json file to identify relationships between datasets and publications.
- title - Canonical title for dataset.
- name - Canonical title for dataset.
- description - Dataset description, if available.
- unique_identifier - Original unique identifier for dataset, normally a DOI if available.
- methodology - Methodology for dataset, if available.
- date - Date when dataset was published, if available.
- coverages - Geographic coverages, if available.
- family_identifier - Internal system ID, roughly captures datasets that have multiple years but are the same dataset. Inconsistently applied, should not be used in analysis.
- mention_list - Array of strings for annotated mentions as identified by human reviewers. Not an exhaustive list of mentions for any given dataset, and only populated for those data sets cited in the phase 1 train-test corpus.

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

ICPSR captured when a given data set was used in analysis within a particular publication, but it did not capture particulars on how that determination was made. To provide better data for participants, we implemented a human content analysis protocol to capture mention text for each data set-publication pair included in our train-test corpus (see [Data Set Mention Annotation Process{.underline}](#_vofr8k96bcvl) below). Since we manually created this data, given limited time and resources, we initially only did this work for data sets that the teams would be using for training and testing in phase 1. In future work, we intend to provide this kind of information for all data sets of interest, and to refine the protocol to capture the exact position in the text of each mention along with the verbatim text.

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

- citation_id - A unique ID for the relationship between one dataset and one publication
- publication_id - Unique ID for a publication which is the same ID for the publication in publications.json
- data_set_id - Unique ID for a dataset which is the same ID for the dataset in the data_sets.json file.
- mention_list - Optional array of strings for alternative references for the dataset in the specific publication (only present in citations included in train-test corpus, and even then, could still be empty).
- score - Confidence score for the dataset being found in the related publication. In ICPSR-specified citations, the score will be 1.0. In model-created files, will depend on the model.

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

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

Data Set Mention Annotation Process

One long-term goal of our efforts in data set detection is to build generalized models that are not overly dependent on use of formal titles of data sets. We aim for models that know of and use the language of discussing and using data to recognize where data is discussed in a particular article and then identify which data sets. The ICPSR data contains many explicit ties between publications and data sets that would have been hard to come by otherwise, but the lack of any indication of which parts of the publication indicated the citation relationship made it difficult to identify the linguistic context within the publication that captured the relationship.

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

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

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

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

The general process:

- each user was assigned a list of citations to code.
- Once the user logged in to the coding tool, they were presented with a list of the coding tasks assigned to them that included a status of each, so they could track which they had already completed, and a link for each to the coding page.

- Once the user loads a particular citation for coding, they are presented with the following coding page, and are asked to follow the coding instructions in the codebook/documentation for the annotation tool ([[https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit]{.underline}](https://docs.google.com/document/d/1xuZL_-z1re6TO3Sv8_9tdFk7z6ovyqTwDVgc1bYO3Ag/edit)):

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

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

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

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

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

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

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

Methods and Fields

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

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

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

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

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

Developing a submission process

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

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

More specifically, we had the following requirements:

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

Building and Submitting a Model

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

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

This work folder contained:

- the script "rcc.sh" and its accompanying configuration "config.sh", that implements all of the basic actions needed to manage docker for a model.
- A set of scaffold files and folders that demonstrate how to hook a model into a docker container, including a Dockerfile with examples of installing OS packages and python packges in a docker container and an example "project" folder with a "code.sh" shell script that is called by default when the docker container is run, pre-configured to call a provided example python file named "project.py".
- A copy of the git repo, for use by the scripts.
- A copy of the dev fold, in the standard data folder structure.

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

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

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

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

Model API

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

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

This data folder has a standard directory structure:

```
data\
|_input\
| |_files\
| |_text\
| |_pdf\
|_output
```

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

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

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

Running a Submitted Model

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

- For each submission, an AWS instance was spun up from a standard image pre-configured to run models built using our submission framework.
- The evaluator connected to the instance and started a screen session, so work would not be disrupted if connection to server was lost.
- The model was downloaded to the server and extracted.
- The submission container was built on the server using the provided Dockerfile and "rcc.sh", and then the container was run over the dev fold to test basic functionality of the container and the model, and to give an estimate of time needed to complete.
- Once the dev fold was successfully processed, "config.sh" was reconfigured to point at the evaluation corpus, and the model was run over the evaluation corpus.
- Once the model completed, standard evaluation Jupyter notebooks in the git submission framework repository were configured to the current projects output and run to generate materials for judges to evaluate the submission.
- Output and results were copied to a central storage area, and the instance used to run the model was terminated.

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

Notes on the Submission Process

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

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

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

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

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

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

Evaluation

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

Phase 1 Evaluation

Mentions, Methods and Fields

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

Data Set Citations

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

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

Phase 2 Evaluation

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

Mentions, Methods and Fields

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

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

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

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

{width="6.5in" height="5.013888888888889in"}

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

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

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

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

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

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

Data Set Citations

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

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

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

Capturing Data References

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

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

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

Finding Related Mentions and Citations

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

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

In our protocol
([[https://docs.google.com/document/d/1Hi13N6gfiRz9nfwCoUQrey8v_ozY7fKHMtHV4GgX2ys/edit#]{.underline}]
(https://docs.google.com/document/d/1Hi13N6gfiRz9nfwCoUQrey8v_ozY7fKHMtHV4GgX2ys/edit))
,
we describe the coding task as “When you are judging data mentions, we want to mark mentions on the right as ”exists“ if they are related to

the data referenced on the left, and make sure to not mark any mentions as "exists" that are not related.”, balanced with “If in doubt, don’t mark a given mention as related.”

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

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

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

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

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

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

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

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

Scoring the Results

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

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

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

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

Discussion

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

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

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

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

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

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

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

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

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

Conclusion

Given the time and resources available, we consider the competition design to have been effective. We got a good number of participants, the resulting models were interesting and some of the solutions were novel and surprisingly effective given their novelty, and discussions after the competition lead to collaborations between pairs of sponsors and participants and collective work on making a gold standard corpus that could be used to develop better models in the future (a great step toward higher quality models). The models also ended up being re-usable as they are, though in a limited scope, and Bundesbank has been able to run them and get output of high enough quality that it is useful to them.

There remains substantial work needed to move this effort forward, however. The next iteration of the competition is tentatively scheduled to begin at the end of 2020, and in this round, we are exploring options for building out a better corpus that combine manual, automated and crowd-sourced means of annotating data. We are working on a more standardized and carefully designed model packaging framework, to facilitate re-use. We are also working on more detailed specifications of model requirements (ability to retrain on data sets of interest without needing a whole new corpus of train-test data, for example). With these changes, we hope to do our part to build on the work of this competition and, next time, start to put in place a framework that will enable us to increase the pace of the innovation we've started.

References

Boudreau, K., & Lakhani, K. (2009). How to Manage Outside Innovation. MIT Sloan Management Review; Cambridge, 50(4), 69–76.

Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & 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. https://doi.org/10.1136/amiajnl–2011–000465

Kalil, T., & Miller, J. (2015, January 13). BUILDING AND USING THE INNOVATION TOOLKIT. Retrieved June 17, 2019, from https://github.com/18F/better-government/wiki/OSTP-Innovation-Toolkit-Memo

Kremer, M., & Williams, H. (2010). Incentivizing Innovation: Adding to the Tool Kit. Innovation Policy and the Economy, 10(1), 1–17. https://doi.org/10.1086/605851

Lewis, S. C., Zamith, R., & 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(1), 34–52. https://doi.org/10.1080/08838151.2012.761702

Mateos-Garcia, J., & Steinmueller, W. E. (2008). The institutions of open source software: Examining the Debian community. Information Economics and Policy, 20(4), 333–344. https://doi.org/DOI: 10.1016/j.infoecopol.2008.06.001

Murray, F., Aghion, P., Dewatripont, M., Kolev, J., & Stern, S. (2016). Of Mice and Academics: Examining the Effect of Openness on Innovation. American Economic Journal: Economic Policy, 8(1), 212–252. https://doi.org/10.1257/pol.20140062

Reardon, S. (2014). Text-mining offers clues to success. Nature News, 509(7501), 410. https://doi.org/10.1038/509410a

Riffe, D., Lacy, S., & Fico, F. G. (2005). Analyzing Media Messages: Using Quantitative Content Analysis in Research, Second Edition (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Sadowski, B. M., Sadowski-Rasters, G., & Duysters, G. (2008). Transition of governance in a mature open software source community: Evidence from the Debian case. Information Economics and Policy, 20(4), 323–332. https://doi.org/DOI: 10.1016/j.infoecopol.2008.05.001

Soboroff, I. M., Ounis, I., Lin, J., & Macdonald, C. (2013). Overview of the TREC–2012 Microblog Track. NIST Special Publication 500–298: The Twenty-First Text REtrieval Conference Proceedings (TREC 2012), 2012, 20. Retrieved from https://www.nist.gov/publications/overview-trec–2012-microblog-track

Tong, A., Diduch, L., Fiscus, J., Haghpanah, Y., Huang, S., Joy, D., … Soboroff, I. (2018). Overview of the NIST 2016 LoReHLT evaluation. Machine Translation, 32(1), 11–30. https://doi.org/10.1007/s10590–017–9200–8

Williams, H. (2012). Innovation Inducement Prizes: Connecting Research to Policy. Journal of Policy Analysis and Management, 31(3), 752–776. https://doi.org/10.1002/pam.21638

Wissler, L., Almashraee, M., Díaz, D. M., & Paschke, A. (2014). The Gold Standard in Corpus Annotation. 5th IEEE Germany Student Conference, IEEE GSC 2014. Presented at the 5th IEEE Germany Student Conference, IEEE GSC 2014, June 26–27, 2014, Passau, Germany. Retrieved from

Wright, B. D. (1983). The Economics of Invention Incentives: Patents, Prizes, and Research Contracts. The American Economic Review, 73(4), 691–707.

Placeholder between chapters

Chapter Break

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)[5] for finding relevant information in the scientific literature. Semantic Scholar extracts meaningful structures in a paper (e.g., images, entities, relationships) and links them to other artifacts when possible (e.g., knowledge bases, GitHub repositories), hence our interest in the rich context competition (RCC). In particular, we participated in the RCC in order to explore methods for extracting and linking datasets used in papers. At the time of this writing, Semantic Scholar comprehensively covers the computer science and biomedical literature, but we plan to expand our coverage in 2019 to other scientific areas, including social sciences.

In the following sections, we describe our approach to the three tasks of the RCC competition: extracting datasets used in publications (Section 2{reference-type="ref" reference="sec:datasets"}), research area prediction (Section 3{reference-type="ref" reference="sec:areas"}) and research method extraction (Section 4{reference-type="ref" reference="sec:methods"}).

Dataset Extraction and Linking {#sec:datasets}

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

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

Provided Data.

The provided knowledge base of known datasets includes approximately 10K datasets used in social science research. Many of the datasets in the knowledge base are specific years or sections of larger surveys, e.g.,

- Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, 1980
- Monitoring the Future: A Continuing Study of the Lifestyles and Values of Youth, 1983
- Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 1996

The high textual similarity between different datasets in the knowledge base makes the linking task more challenging.

In the first phase of the RCC, organizers provided participants with 5K papers, partially annotated with dataset usage to serve as training data. For each paper, the full text, metadata, and PDF file were provided. For each annotation, the paper ID and the corresponding dataset ID in the knowledge base were provided. For most annotations, the textual mentions in the paper were also provided, but the position of the mention in the paper text was not specified. This means that for a mention that appears multiple times in a paper, it is ambiguous which of these mentions is actually the reference to the dataset that has been labeled. In order to actually label the text in the paper, we need to search the paper for the annotated mention, and if the mention appears multiple times, it is not clear which of these are valid examples of dataset usage in a paper. Approximately 10% of the datasets in the knowledge base were linked one or more times in the provided corpus of 5K papers.

{width="13cm"}

We provide a high-level overview of our approach in Figure [fig:datasets]{reference-type="ref" reference="fig:datasets"}. First, we use a named entity recognition (NER) model to predict dataset mentions. For each mention, we generate a list of candidate datasets from the knowledge base. We also developed a rule based extraction system which searches for dataset mentions seen in the training set, adding the corresponding dataset IDs in the training set annotations as candidates. We then use a binary classifier to predict which of these candidates is a correct dataset extraction.

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

Mention and Candidate Generation.

We first constructed a set of rule based candidate citations by exact string matching mentions and dataset names from the provided knowledge base. We found this to have high recall on the provided development fold and our own development fold that we created. 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.

To address this limitation, we developed an NER model to predict additional dataset mentions. For NER, we use a bi-LSTM model with a CRF decoding layer, similar to [@Peters2018DEEPCW], and implemented using the AllenNLP framework.[6] In order to train the NER model, we automatically generate mention labels by string matching mentions in the provided annotations against the full text of a paper. This results in noisy labeled data, because it was not possible to find all correct mentions this way (e.g., some dataset mentions were not annotated), and the same string can appear multiple times in the paper, while only some are correct examples of dataset usage.

We limit the percentage of negative examples (i.e., sentences with no mentions) used in training to 50%, and use 40 words as the maximum sentence length. We use 50-dimensional Glove word embeddings [@Pennington2014GloveGV], 16-dimensional character embeddings with 64 CNN filters of sizes (2, 3, 4). The CNN character encoder outputs 128-dimensional vectors. We optimize model parameters using ADAM [@Kingma2014AdamAM] with a learning rate of 0.001.

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

Candidate Linking.

The linking model takes as input a dataset mention, its context, and one of the candidate datasets in the knowledge base, and outputs a binary label. We use a gradient boosted trees classifier using the XGBoost implementation.[7] We use the following features: prior probability of entity, prior probability of entity given mention, prior probability of mention given entity, whether a year appears in the mention context and in the dataset title, mention length, mention sentence length, whether the mention is an acronym, estimated section title of the mention, overlap between mention context and dataset keywords provided in the knowledge base, and the TF-IDF weighted token overlap. We note that it is possible to predict zero, one or multiple dataset IDs for the same mention.

Results.

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

We also experimented with adding ELMo embeddings [@Peters2018DEEPCW], 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.

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

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

[[tab:ner_results]]{#tab:ner_results label="tab:ner_results"}

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

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

[[tab:test_results]]{#tab:test_results label="tab:test_results"}

We report the end-to-end performance of our approach (on the development set provided by the organizers in the first phase) in Table [tab:e2e_results]{reference-type="ref" reference="tab:e2e_results"}. 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 [tab:e2e_results]{reference-type="ref" reference="tab:e2e_results"} are not indicative of test set performance. End to end performance from our phase 2 submission can be seen in Table [tab:test_results]{reference-type="ref" reference="tab:test_results"}. This performance is reflective of our focus on the linking component of this task. Aside from the competition development set, we also used a random portion of the training set as an additional development set. The initial model only uses a dataset frequency feature, which gives a baseline performance of 38.4 F1. Adding p(d | m) and p(m | d), which are the probability of entity given mention and probability of mention given entity improves the performance ($\Delta = 2.3$ F1). Year matching helps disambiguate between different datasets in the same series, which was found to be a major source of errors in earlier models ($\Delta = 2.8$ F1). Aggregating mentions for a given dataset, adding mention and sentence length features, adding an is acronym feature, and further hyper-parameter tuning improve the results ($\Delta = 12.5$ F1). Adding examples in the development set while training the model results in further improvements ($\Delta = 2.8$ F1). Finally, adding the NER-based mentions significantly improves recall at the cost of lower precision, with a positive net effect on F1 score ($\Delta = 0.7$ F1).

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

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

[[tab:e2e_results]]{#tab:e2e_results label="tab:e2e_results"}

Research Area Prediction {#sec:areas}

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) [@Shen2018AWS], we decided to use a subset of the fields of study in MAG as labels. In particular, we included all fields related to social science or papers from the provided training corpus. However, since the abstract and full text of papers are not provided in MAG, we only use the paper titles for training our model. The training data we ended up with included approximately 75K paper titles along with their fields of study as specified in two levels of the MAG hierarchy. We held out about 10% of the titles for development data. The coarse level (L0) has 7 fields while the more granular one (L1) has 32. Fields associated with less than 100 papers were excluded.

Methods.

For each level, we trained a bi-directional LSTM which reads the paper title and predicts one of the fields in this level. We additionally incorporate ELMo embeddings [@Peters2018DEEPCW] 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 it exceeds a certain threshold. It takes approximately 1.5 and 3.5 hours for the L0 and L1 classifiers to converge, respectively.

Results.

To select a model, we performed a 100 trial random search across model hyper-parameters, evaluated on a held out development set of papers from the Microsoft Academic Graph. Our final model contained 512 hidden dimensions, 2 layers and 0.5 dropout prior to classification. The top performing classifier achieved 84.4% accuracy on our development set on L0 fields, and 65.2% accuracy on our development set on L1 fields.

Research Method Extraction {#sec:methods}

Data.

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

Methods.

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

Results.

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 a 95% precision and yield of 1.5 methods per paper on the manually inspected subset of papers.

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, 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.

Future Work

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. For research area prediction, it would help to include signals beyond just the paper title for predicting the field of study. The difficulty here is finding labeled training data that includes richer signals like abstract text and paper keywords. For method prediction, further exploration of using 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.

Acknowledgments {#acknowledgments .unnumbered}

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 (Siri) Shen for helping us use the MAG data.

Non technical overview

Our approach for retrieving datasets is to generate our own questions about dataset names and use a machine learning technique to train the model for solving question answering task. In other words, questions suitable for finding dataset names such as “What is the dataset used in this paper?,” are generated and the question answering model is trained to find the answers to those questions from the papers. Furthermore, the resulting answers from the model are filtered by types of each word. For example, if an answer contains words with organization or agency types, then this answer is likely to include the actual dataset names.

For the research fields retrieval, we first crawled Wikipedia articles that correspond to the list of research fields. Then, we retrieved the research fields of the papers 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 paper and the article “economic history” is high enough, it is determined that the paper belongs to a research field “economic history.” For the similarity measurement, the TF-IDF similarity is used, which is the similarity measurement based on the term frequency and document frequency.

For the research methods retrieval, we train a model that recognizes named entities through a machine learning technique. More specifically, we considered the research methods as named entities and train a model to retrieve the named entities. For example, for a research method called “bivariate analysis”, this research method is considered as a named entity by the trained model and therefore, retrieved by the model.

Literature Review

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

What did you do

In this section, we will explain about the models we used for datasets retrieval, research fields retrieval, and research methods retrieval.

Datasets Retrieval

Our approach to solving the dataset retrieval task is reading comprehension (RC) with query generation and entity typing. An RC model is applied to the given publications with our own query generation module. Then, the result from the RC model is filtered with an entity typing module. Figure 1 shows our overall approach for dataset retrieval. In following subsections RC model, query generation, and entity typing are explained in detail.

image

Figure 1: Overall architecture for dataset retrieval

Document QA

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

The RC model used in this work is Document QA [@clark2017simple]. It uses Bi-GRU, bi-attention, and self-attention mechanism. In addition, Document QA performs a paragraph selection that pre-filters and selects the *k* most relevant paragraphs through TF-IDF similarity between the query and paragraphs. We observed that datasets are usually mentioned together in some specific paragraphs of the publications. Therefore, this model is appropriate for this task thanks to its paragraph selection stage.

Query generation module

In order to apply an RC model, such as Document QA, to the dataset retrieval task, queries that are suitable for finding the datasets are required. However, defining a general query for retrieving datasets is difficult, since the dataset mentions appear in various forms like surveys, datasets, or studies. Therefore, we devised a query generation module with some important query terms to generate multiple specific queries instead of one general query.

To generate important query terms, we used a query generation model that creates queries given answers proposed by [@yuan2017machine]. Thanks to this model, we could obtain a list of queries to retrieve datasets from the training set. After that, we extracted query terms that are frequent in the list of queries and at the same time are not frequent in non-dataset-mention sentences.

Because of this, these query terms have discrimination power for retrieving dataset mentions since 1) queries are generated to extract mentions and 2) the query terms do not appear in the sentences without dataset mentions.

This list of query terms is used to generate a general query concatenating query terms. This query is used for the paragraph selection stage of Document QA, as shown in Figure 1. After this stage, the query generation module generates queries for each paragraph by string matching, in order to create specific queries for each paragraph.

Ultra-Fine Entity Typing

Ultra-Fine Entity Typing [@Choi:2018:ACL] can predict a set of free-from phrases like *criminal* or *skyscraper* given a sentence with an entity mention. For example, in the sentence: *Bob robbed John and he was arrested shortly afterward*, Bob is of type *criminal*. In our task, candidate answers proposed by Document QA and their context are input to Ultra-Fine Entity Typing. Although this system can predict 10k different entity types in which *dataset* is included, after a few experiments we observed that most of the dataset names are recognized as some specific entity types such as *organization* and *agency*. Since these entity types are consistent, we decided that this could be a feature for our candidate answer classifier.

Candidate Answer Classifier

Using the score given by the RC model for each candidate answer and the entity types given by Ultra-Fine Entity Typing for each candidate answer, a neural network classifier that filters the candidate answers of Document QA was used. We discovered that a candidate answer with a high score given by Document QA and whose entity type is *organization* or something similar is considerably likely to be a correct dataset name. Due to this pattern, we were able to create neural network classifier to filter out candidate answers.

The classifier has the following architecture:

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

The training set consists of 25172 examples and the test set of 6293 examples. Adam optimizer was used and cross entropy was used as loss function.

Research Fields Retrieval

Our approach to obtaining the research fields is based on TF-IDF similarity with Wikipedia articles. First, a set of Wikipedia articles about different research fields using the library MediaWiki for Python was obtained. The list of research fields provided the Coleridge Initiative for the Rich

Context Competition was used to crawl Wikipedia. This list has three levels of hierarchy as the example in Figure 2.

Figure 2: Research fields hierarchy

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 [@scikit-learn], a TF-IDF matrix of all the publications and Wikipedia articles of research fields were computed and the research field and all its superior nodes in the hierarchy associated with the most similar article were returned along with the similarity in the range [0,1]. The overall architecture can be seen in Figure 3.

Figure 3: Overall architecture for research fields retrieval

Research Methods Retrieval

For the research methods retrieval task, we modeled it as an named-entity recognition (NER) problem. Research methods are considered to be a named entities and because of this, they can be tagged as research method label (RS) instead of common NER labels such as: *location*, *people*, etc. Figure 4 shows the main architecture of the model proposed by [@lample2016neural] and used in this task.

Figure 4: Paragraph selection for DocQA in research method retrieval

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

In this task, research method phrases which appeared in the training set were marked. Then, we represented the data in CoNLL 2003 format [@tjong2003introduction], using IOB tag (Inside, Outside, Beginning). Every token is labeled as B-RS if the token is the beginning of a research method, I-RS if it is inside a research method but not the first token, or O if otherwise. We used this type of data to train the model which could detect research methods in publications.

What worked and what didn’t

We tried different ideas to extract dataset names. Firstly, we tried to extract the dataset name using some hand-crafted queries in the QA model. But 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 some generalized queries. However, we found that each of the resulting clusters did not reflect generalized queries. Hence, we had to create specific queries for each publication as explained in the previous section.

We also tried to use the section names as a feature of the paragraph selection part in the Document QA. However, the use of section name has rather degraded the overall performance. In our analysis, this seems to be due to the noise that occurred when extracting the section name, since we relied on some heuristics to extract the section names.

The use of entity typing worked well to remove the wrong candidate answers proposed by the reading comprehension model. Thanks to this filtering by entity types, we were able to improve the recall using the query generation module without sacrificing the precision.

Our approach to retrieve research fields worked well as will be shown in the next section.

Finally, our first idea to retrieve research methods was based on identifying their context words by using the frequency of those words. However, this approach did not achieve good results due to 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, where we consider each research method that appeared in a paper as a named-entity. By modeling the problem in this way, we can use existing NER models to extract research methods from papers. However, this approach also performed poorly. We found that the dataset we used for training was not appropriate for this task. We will provide an in-depth analysis of this problem in the next section.

Summary of your results and caveats

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

Datasets Retrieval

To analyze the effects of the query generation module and entity typing module, we performed analyses on 100 phase 1 dev set with 3 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 5 shows the results from 3 publications of phase 1 dev set with Document QA only. Compared to the other settings, Document QA only setting retrieves answers (dataset mentions) with high quality. However, the number of retrieved answers is notably small. For example, the result from *153.txt* publication was empty as in Figure 5. In fact, our model using this setting can retrieve only 260 answers (predictions) from 100 publications of phase 1 dev set.

Figure 5: Results from Document QA only

These results with fewer answers were expected, due to the difficulty of defining general queries as explained in section *Question Generation Module*. Without a query generation module, our query was not representative enough to retrieve various forms and types of the dataset mentions.

Document QA + query generation module

Figure 6 shows the results from 3 publications of phase 1 dev set with Document QA and query generation module. Because of the latter, our dataset retrieval model could retrieve a large number of answers. For example, the result from *153.txt* publication contains a large number of answers with correct answers such as *financial services FDI data* or *Micro Batabase Direct investment*. Therefore, we believe that the query generation module improves recall of the entire dataset retrieval model. Actually, our model using this setting can retrieve more than 2,000 answers (predictions) from 100 publications of phase 1 dev set.

However, compared to the Document QA only setting, there is a considerable number of noise. For example, in Figure 6, *empirical, Table 1, Section 4* and etc., are not dataset mentions.

Figure 6: Results from Document QA + query generation module

We believed that the reason of these noises is the several query terms potentially retrieve wrong answers. For example, we have a query term "*study*" to retrieve dataset mentions such as "*ANES 1952 Time Series Study*". However, this term can also retrieve noises such as "*empirical study*". These kinds of query terms are still needed to retrieve various forms and types of dataset mentions but clearly generate some noises.

Document QA + query generation module + entity typing module

Figure 7 shows the results from 3 publications of phase 1 dev set with Document QA, query generation module, and entity typing module. Thanks to the entity typing module, we can see that most of the noises from the query generation module have disappeared. Although a few right answers such as "*FDI data*" was filtered out and a few wrong answers such as "*4.2.1 Micro Data*" was not, overall precision is adequately improved by entity typing module. In addition, our model in this setting could retrieve 526 answers (predictions) from 100 publications of phase 1 dev set.

Figure 7: Results from Document QA + query generation module + entity typing module

Research Fields Retrieval

We randomly selected 20 publications from the training set of phase 1, since our model does not require any training. The model was able to correctly predict 11. The strongest point is that the model is able to predict research fields which are significantly specific such as *Home health nursing management*. Among the weak points of the model, it has problems when two research fields are similar or share subtopics. Moreover, sometimes it fails due to the fact that it tries to retrieve excessively specific fields while more general ones would be suitable.

Research Methods Retrieval

20 random publications were selected from the training set of phase 2 and labeled. Our result is not as expected. The model is able to find proper research methods for 12 publications out of 20. For example, the model detects one of the research methods appeared in publication with id 15359 which is *Factor analysis*. However, the results contain a notable amount of noise. For example, the document with id 10751, the model retrieves several wrong answers like *Reviews describe*, *Composite materials*, *Detailed databases*, etc. After analyzing this result, we found that the dataset we use for training is not appropriate for this task. For example, *Reliability* and *Independent variables* are marked as research methods, but actually, they are not.

Lessons learned and what would you do differently

After the completion of this project, we realized that some steps could have been in a different way and led to better results. For example, we focused a lot on the model creation, however, we think that we should have spent more time on the analysis of the dataset to extract all its potential and search for additional datasets since some of the provided datasets contain noise.

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 modeling dataset retrieval as a QA task was right. However, at some point during the project, we should have changed it to another model with a state of the art performance.

Also, in the QA model, we are currently using symbolic queries. But since we are generating our own queries, we could define and generate queries with a distributed representation. It would be more generic and model-matching queries. Furthermore, for research fields, we should have tried other ranking methods like BM25, a ranking function used by search engines whose performance is better than TF-IDF.

Finally, for research methods, because of the noise in the dataset, supervised NER could not achieve the desired performance, so we should have used unsupervised NER to avoid that noise.

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 have a hypothesis that datasets depend on research fields and vice versa. For example, in the research field *Question Answering*, a subfield of *Natural Language Processing* and *Computer Science*, the most commonly used dataset is SQuAD [@rajpurkar2016squad]. Therefore, according to our hypothesis, two publications using SQuAD are presumably to be in the same field, *Question Answering*. Based on this hypothesis, we intend to build hierarchical clusters of publications with the same research field. This way, a cluster will have publications with the same research field and similar datasets. As an example, the QA cluster will have papers about QA and those papers will use similar datasets like SQuAD and TriviaQA [@joshi2017triviaqa]. With these clusters, the system will be able to recommend datasets to data users. For example, if a publication is in the *Question Answering* field, the proposed system would be able to recommend the authors SQuAD and TriviaQA. Moreover, it would be able to recommend to data producers fields with a lack of datasets.

In addition, we also need to improve the performance of the models we built. For example, since we used a pre-trained model in Document QA we think we could not exploit the whole potential of this system, so we would like to train our own model using a training set of publications.

Appendix

Description of your 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

Placeholder between chapters

Chapter Break

Introduction

Scientists and analysts often face the problem of finding interesting research datasets and identifying who else used the data, in which research fields, and how the data has been analyzed from a methodological perspective. To address these problems, the Coleridge Initiative organized the Rich Context Competition[1]. The competition invited international research teams to develop text analysis and machine learning tools that can discover relationships between research datasets, methods, and fields in scientific literature. The competition took place between October 2018 and February 2019 and included two phases[2]. The first phase was open for all teams which have submitted a letter of intent. Teams are then provided with a corpus of social science publications to develop and train machine learning algorithms for automatic research dataset, methods and field detection and linking. More concretely, one major subtask consisted of linking dataset mentions to a given set of around 10,000 dataset descriptions from the ICPSR's research data index.[3] Only the best four teams from the first phase are invited to the second phase of the competition and asked to discover research datasets, methods, and fields in a larger corpus of social science publications. All submitted algorithms have to be made publicly available as open source tools. With this document, we (team RCC–5) aim to fulfill another requirement, i.e., the documentation and summary of the developed approach including data pre-processing, algorithms, and software.

General Approach and Software Components

One of the central tasks in the RCC is the extraction of dataset mentions from text. Nevertheless, we considered the methods and fields discovery equally important. To this end, we decided to follow a module-based approach and developed tools that can be used separately but also as parts of a data processing pipeline. Figure [figure:pipeline] shows an overview of the software modules developed for the RCC competition, including their dependencies. Here, the upper three modules (gray) describe the pre-processing steps (cf. Section [sec:prepro]). The lower four modules (blue) are used to generate the output in a pre-specified format. The pre-processing step consists of extracting metadata and pure text from PDF documents. The extraction itself is done using the Cermine Tool[4] which returns a Journal Article Tag Suite[5] XML document. Then, in a second step, text, metadata and references are extracted. The output of the pre-processing is then used by the software modules responsible for tackling the individual sub-tasks, i.e., discovering research datasets (cf. Section [sec:dataset-extraction]), methods (cf. Section [section:research_method_extraction]) and fields (cf. Section [section:field_classification]). Section [sec:techdoc] provides the technical details of the modules, i.e., input, output, and how to run the modules.

First Phase Feedback

After the first phase, each team received feedback from the organizers of the RCC. The feedback is twofold and consists of a quantitative and qualitative evaluation. Unfortunately, our team did not perform very well regarding precision and recall. In contrast to this, our approach has been found convincing regarding the quality of results. The qualitative feedback result from a random sample of ten documents that are given to four judges. Judges are then asked to manually extract dataset mentions and calculate the overlap between their dataset extractions and the output of our algorithm. Other factors that judges took into consideration are specificity, uniqueness and multiple occurrences of dataset mentions. As for the extraction of research methods and fields no ground truth has been provided, these tasks were evaluated against the judges’ expert knowledge. Similarly to the extraction of dataset mentions, specificity and uniqueness have been considered for these two tasks. The feedback our team received acknowledged the fact that no ground truth has been provided and our efforts regarding the extraction of research methods and fields.

Data and Pre-processing

This section describes the data provided by the organizers of the RCC, the external data sources we used as well as our pre-processing steps.

The RCC Corpus

For the first phase, the data provided by the organizers consisted of 5,000 publications. Additionally, a development fold of 100 plain text publications, their metadata, a list of datasets of interest (including all datasets that were explicitly referenced in the curated corpus) were given. The list of datasets should not be considered complete as there could be additional datasets mentioned in these publications. The organizers also provided examples of social science research methods and fields vocabularies in term of SAGE Publications research field and method vocabularies. In the second phase of the competition, an additional set of 5,000 publications from the social sciences has been provided.

External Data Sources

For developing our algorithms, we also utilized two external data sources. For the discovery of research methods and fields, we resort to data from Social Science Open Access Repository6. SSOAR is maintained at GESIS – Leibniz Institute for the Social Sciences collects and archives literature of relevance to the social sciences. In SSOAR, full texts are indexed using controlled social science vocabulary (Thesaurus[7], Classification[8]) and are assigned rich metadata. SSOAR offers documents in various languages. The corpus of English language publications that can be used for purposes of the competition consists of a total of 13,175 documents. All SSOAR documents can be accessed through the OAI-PMH[9] interface. Another external source that we used for discovery of 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.

Pre-processing

Although the organizers of the RCC, offered plain texts for the publication, we decided to build our own pre-process pipeline. The pipeline uses the Cermine Tool to extract information from PDF documents. The main benefit of using this tool is the structured metadata output including better disambiguation of sections and paragraphs in the publications. The output XML file uses the Journal Article Tag Suite[10]. For the competition, there are only two interesting elements of the Jats XML format, i.e., <front> and <body>. The <front> element contains the metadata of the publication, whereas the <body> contains the publication text. Another advantage of Cermine is that the hyphenation and segmentation of paragraphs are carried out automatically. As a last step of the pre-processing, we remove all linebreaks from the publication text and output a list of metadata fields and values as shown in Table [tab:example-paragraph] for each publication paragraph.

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

[tab:example-paragraph]

Dataset Extraction

Task Description

In scientific literature, datasets are specified to indicate, e.g., the data on which a analysis is performed, a certain finding or a claim is based on. In this competition, we focus on (i) extracting and (ii) linking datasets mention from social science publications to a list of given dataset references. Identifying dataset mention in literature is a challenging problem due to the lack of an established style of citing datasets. Furthermore, in many research publication, a correct citation of datasets is entirely missing (Boland et al., 2012). The following two sentences exemplify the problem.
Example 1: *P-values are reported for the one-tail paired t-test on* Allbus* (dataset mention) and *ISSP* (dataset mention).*
Example 2: *We used* WHO data from 2001* (dataset mention) to estimate the spreading degree of AIDS in Uganda.*
We treat the problem of detecting dataset mentions in full-text as a Named Entity Recognition (NER) task.

Formal problem definition

Let *D* denote a set of existing datasets *d* and the knowledgebase *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

With our method, we focus on the extraction of dataset mentions in the body of the full-text of scientific publications. We recognize three types a dataset can be mentioned: (i) The full name of a dataset like ”National Health and Nutrition Examination Survey“, (ii) an abbreviation (”NHaNES“) or (iii) a vague reference, e.g., ”the monthly statistic“. By each of these varieties, the NER task faces particular challenges. For the first type, the used dataset name can vary in different publications. Where one publication cites the dataset with ”National Health and Nutrition Examination Survey“ the other could use the words ”Health and Nutrition Survey“. In a case where abbreviations are used a disambiguation problem occurs, e.g., in ”WHO data“. WHO may describe the World Health Organization or the White House Office. The biggest challenge is again the lack of a precise gold standard that can be used to train a classifier. In the following we describe how we have dealt with this lack of ground truth data.

Phase one approach

The challenge of missing ground truth data is the main problem to handle during this competition. To this end, supervised learning methods for dataset mentions extraction from text are not directly applicable. To overcome this limitation, we resort to the provided list of dataset mentions and publication pairs and re-annotate the particular sentences in the publication text. This re-annotation is then used to train Spacy’s neural network based NER model[11]. We created a holdout set of 1000 publications and a training set of size 4000. We train our model using publication paragraphs as training samples. In the training set, 0.45 percent of the paragraphs contained mentions. For each positive training example, we added a negative example that does not contain dataset mentions and is sampled at random. 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: strict precision and recall, and partial precision and recall. While the former are standard evaluation metrics, the latter are their relaxed variants in which the degree to which dataset mentions have to match can vary. Consider the following example of a partial match: “National Health and Nutrition Examination Survey” is the extracted dataset mention whereas, “National Health and Nutrition Examination Survey (NHANES)” represents the true dataset mention.

Metric	Value
Partial Precision	0.93
Partial Recall	0.95
Strict Precision	0.80
Strict Recall	0.81

[table:dataset-mention-eval]

Table [table:dataset-mention-eval] show the results of the dataset mention extraction on the holdout set. The model is able to achieve high strict precision and recall values. As expected, the results are even better for the partial version of the metrics. But, this version indicates that even if we are not able to exactly match the dataset mention in text, we can find the right context with very high precision at least.

Phase two approach

In the second phase of the competition additional 5,000 publications have been provided. We extended our approach to consider the list with dataset names supplied by the organizers and re-annotated the complete corpus of 15.000 publication in the same manner as in phase one to obtain training data. This time we split the data in 80% for training and 20% for test.

Evaluation

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

Metric	Value
Partial Precision	0.51
Partial Recall	0.90
Strict Precision	0.49
Strict Recall	0.87

[table:dataset-mention-eval-phase-two]

Research Method Extraction

Task Description

Inspired by a recent work of Nasar et al. (Nasar et al., 2018), we define a list of basic entity types that give key-insights into scholarly publications. We adapted the list of semantic entity types to the domain of the social sciences with a focus on *research methods*, but also including related entity types such as *Theory, Model, Measurement, Tool, Performance*. We suspect that the division into semantic types might be helpful to find *research methods*, because related semantic entities types might provide clues or might be directly related to the research method itself. For instance, in order to realize a certain research objective, an experiment is instrumented where a specific combination of *methods* is applied to a *data set* that might be intellectual or *software*, thus achieving a specific

performance and result in that context.
Example: *P-values* (measurement) are reported for the *one-tail paired t-test* (method) on *Allbus* (dataset) and *ISSP* (dataset).

Formal problem definition

Let *E* denote a set of entities. The Named Entity Recognition and Linking task consists of (i) identifying entity mentions *m* in a sentence and, (ii) linking them, when possible, to a reference knowledge base *K* (i.e, the SAGE Thesaurus[12]) and (iii) assigning a type to the entity, e.g., *research method*, selected from a set of given types. Given a textual named entity mention *m* along with the unstructured text in which it appears, the goal is to produce a mapping from the mention *m* to its referent real world entity *e* in *K*.

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[13]. With respect to entity linking, another challenge is detecting name variations, since entities can be referred to in many different ways. Semantically similar words, synonyms or related words, which might be lexically or syntactically different, are often not listed in the knowledge base (e.g., the lack of certain terms like 'questioning' but not 'questionnaire'). This problem of automatically detecting these relationships is generally known as **linking problem**. Note that part of this problem also results from PDF-to-text conversion which is error-prone. Dealing with incomplete knowledge bases, i.e. **handling of out of vocabulary (OOV) items**, is also a major issue, since knowledge bases are often not exhaustive enough and do not cover specific terms or novel concepts from recent research. Last but not least, the combination of different semantic types gives a more coherent picture of a research article. We hypothesize that such information would be helpful and results in an insightful co-occurrence statistics, and provides additional detail directly related to entity resolution, and finally helps to assess the **relevance of terms** by means of a score.

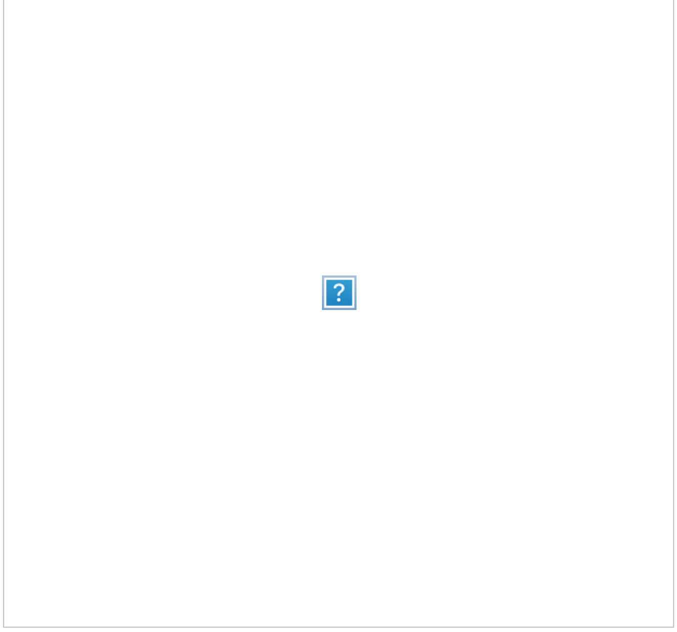
Our Approach - Overview

Our context-aware framework builds on Stanford's CoreNLP and Named Entity Recognition System[14]. The information extraction process follows the workflow depicted in Figure [figure:pipeline], using separate modules for pre-processing, classification, linking and term filtering.

We envision the task of finding entities in scientific publications as a sequence labeling problem, where each input word is classified as being of a dedicated semantic type or not. In order to handle entities related to our domain, we train a novel machine learning classifier with major semantic classes, 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 SAGE dictionary of Social Research Methods (Lewis-Beck et al., 2003), thus putting a special emphasis on the social sciences. The linking component addresses the synonymy problem and matches an entity despite name variations such as spelling variations. Finally, term filtering is carried out based on a termhood and unithood, while scoring is achieved by calculating a relevance score based on TF-IDF (cf. Section [para:relscore]).

Our research experiments are based on the repository for the Social Sciences SSOAR as well as the train and test data of the Rich Context Competition corpus[15]. Our work extends previous work on this topic (cf. (Eckle-Kohler et al., 2013)) in various ways: First, we do not limit our study to abstracts, but use the entire fulltext. Second, we focus on a broader range of semantic classes, i.e. *Research Method*, *Research Theory*, *Research Tool* and *Research Measurement*, tackling also the problem of identifying novel entities.

Overview of the entity extraction pipeline



[pipeline]

Distributed Semantic Models

For domain adaptation, we integrate further background knowledge. We use vector embeddings of words trained on additional corpora and which serve as input features to the 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 from the scientific ACL ARC[16] using Gensim with the skip gram model (cf. (Mikolov et al., 2013)) and a pre-clustering algorithm[17]. A summary of the size of the unlabeled English data used for training word embeddings can be found in Table [tab:UnlabeledData].

Corpus	Articles	Documents/Tokens
ACL Corpus	22,878	806,791/2.5 GB

Features

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

Type	Features
Token unigrams	$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, \ldots$
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’)[18]. However, since the thesaurus is not exhaustive and covers only the top-level concepts related to social science methods, our aim was to extend it by automatically extracting further terms from domain-specific texts, in particular from the Social Science Open Access Repository. More concretely, we carried out the following steps to extend SAGE as an off-line step. For step 2 and 3, candidate terms have been extracted by our pipeline for the entire SSOAR corpus.

1. Assignment of semantic types to concepts (manual)
2. Extracting terms variants such as abbreviations, synonyms, related terms from SSOAR (semi-automatic)
3. Computation of Term and Document Frequency Scores for SSOAR (automatic)

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

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

Computation of Term and Document Frequency Scores

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

[tab:SAGET] 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

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

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

Preliminary Results

Our method has been tested on 100 fulltext papers from SSOAR and 10 documents from the Rich Context Competition (RCC), all randomly selected from hold out corpora. In our experiments on SSOAR Social Science publications, we compared results to the given metadata information. The main finding was that while most entities from the SAGE thesaurus could be extracted and linked reliably (e.g., ’Paired t-test’), they could not be easily mapped to the SSOAR metadata terms, which consist of only a few abstract classes (e.g., ’quantitative analysis’). Furthermore, our tool was tested by the RCC organizer, were the judges reviewed 10 random publications and generated qualitative scores for each document.

Conclusion and Future Work

We plan to carry out a more detailed evaluation on fulltext scholarly publications and assess the impact of different features used in the ML model, including background resources such as embeddings and dictionaries.

Research Field Classification

Task Description

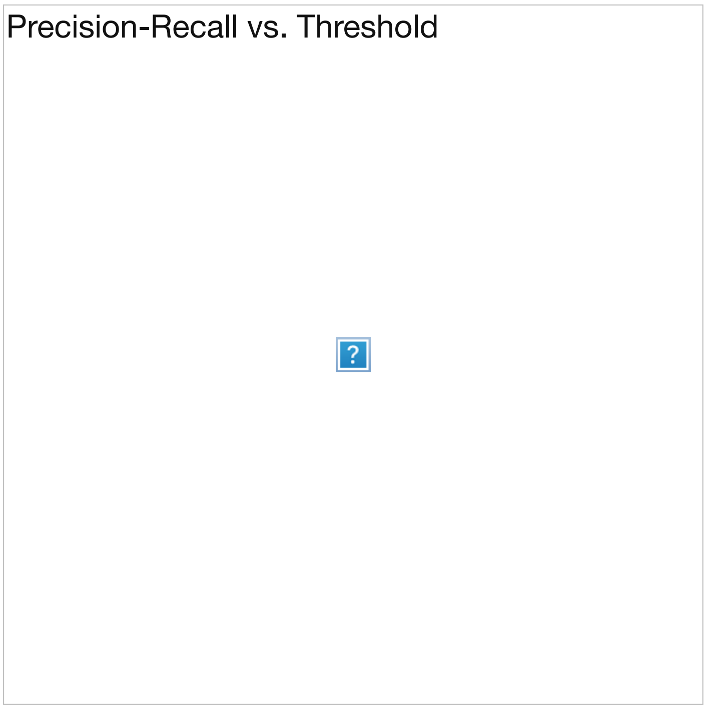
The goal of this task is to identify the research fields covered in social science publications. The RCC data does not provide a gold standard —annotated training data— for that task. To this end, we decided to train a classifier using annotated data from SSOAR. In this way, our interpretation of the task is to select one or more labels from a given set of labels for each publication. This approach is known as a mulit-label classification. In our case, a label represents a research field.

Our approach - Overview

Due to the unequal distribution of labels in the dataset, we need to guaranty enough training data for each label. We selected only labels with frequency over 300 for training the model which results in a total of 44 labels representing research fields. We decided to train a classification model based on the fasttext framework (Joulin et al., 2017). To train our model we resort to the abstracts of the publication, as this approach worked better than using the full-texts.

Evaluation

Figure [fig:results_fasttext] shows the performance of the model regarding various evaluation metrics for different thresholds. A label is assigned to a publication if the model outputs a probability for the label above the defined threshold. In multi-label classification, this allows us to evaluate our model from different perspectives.


Precision-Recall vs. Threshold

Technical Documentation

The project contains the following modules listed in the order in which they are excecuted.

Pre-processing

PDF Text Extraction

Module name :

Cermine_NlmJat_extractor
Function:

Converts each PDF files of a given folder to JATS XML Format. Each input PDF File is transformed to one XML File.
Bash function call:

```
java -jar target/cermineXMLextraction-1.0.0-jar-with-dependencies.jar
```

Parameter (2):

```
-s <source folder>\
-t <target folder>
```

Returns:

XML Files in JATS XML Format.
Build:

This is a Java program using Maven build tool.
build call:

mvn install

Extraction of text from JATS XML

Module name :

preprocess-rcc-data
Function:

Transform text from JATX XML Format into a JSON File containing a list of textfields with essential metadata for each JATS XML file of a given folder.
Bash function call:

```
`python3  ./jats_text_extractor.py `
```

Parameter (4):

```
<source folder>
<target folder>
<limiting number of files to transform (-1: all)>
<number of cores to use for multiprocessing (-1: all)>
```

Returns:

A JSON File for each given XML File in the source folder

Metadata Extraction

Module name :

preprocess-rcc-data
Function:

Extracts structured metadata and references from all JATS XML files in a given folder into two Files. One containing the metadata from all Publications in JATS XML files and one containing all references from the JATS XML Files. The target file format is JSON.
Bash function call:

```
python3  ./jats_metadata_extractor.py
```

Parameter (3):

```
<source folder>
<target filename for metadata>
<target filename for references>
```

Returns:

Two JSON files containing metadata and references from all XML Files

Dataset Mentions

Dataset Mention Extraction

Module:

dataset-mention-extraction
Function:

Extract dataset mentions from all JSON Files from a given folder with a given spacy model.
Bash function call:

```
python3  ./predict_mentions.py
```

Parameter (4):

```
<source folder>
<name of spacy model folder>
<target filename rcc-output>
<target filename internal format>
```

Returns:

Two JSON files containing the found dataset mentions in all given JSON Files. One in RCC defined output. One in Internal format including the senctence the dataset mention occures.
Train:

For training we submit a jupyter notebook with all needed code. *Train_spacy_ner_prod.ipynb*
Build (For training only):

Install english spacy language model. This can be done with ‘python -m spacy download en‘.

Dataset Linking (only Phase 1)

Module:

dataset-prediction
Function:

Links dataset mentions given a JSON file in internal format to datasets listed in a given JSON File.
Bash function call:

```
python3  ./retrieve.py
```

Parameter (3):

```
<JSON filename of extracted mentions>
<JSON filename of dataset list to match>
<output filename for dataset citations>
```

Returns:

JSON file in the format defined by the competition containing information about links between publications and datasets.

Research Method Extraction

Module name :

research-method-extractor
Function:

Extracts research method terms from JSON files with text information from publications.
Bash function call:

```
java -jar target/gesisents-0.1-jar-with-dependencies.jar
```

Parameter (3):

```
<source folder>
<target file name>
<Limit to reduce the number of processed files (-1:all)>
```

Returns:

A JSON file in the format defined by the competition containing information about publications and research methods. **Build:**

This is a Java program using Maven build tool.
build call:

mvn install

Research Field Classifier

Module name :

research-field-detector
Function:

Classifies given abstracts with classoz Labels
Bash function call:

```
python3  ./fasttext_predictor.py
```

Parameter (4):

```
<filename of JSON file with abstracts>
<filename of fasttext model>
<filename of label dictionary in JSON>
<target filename labels in >
```

Returns:

A JSON file in the format defined by the competition containing information about publications and research fields.

Acknowledgments

We would like to thank GESIS for giving us the time and resources to participate in the competition.

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

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

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

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

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

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

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*, April 2017, pp. 427–431. Association for Computational Linguistics.

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

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

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

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

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

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

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

[1] https://coleridgeinitiative.org/richcontextcompetition

[2] https://coleridgeinitiative.org/richcontextcompetition#competitionschedule

[3] https://www.icpsr.umich.edu/index.html

[4] https://github.com/CeON/CERMINE

[5] https://jats.nlm.nih.gov

[6] https://www.gesis.org/ssoar/home

[7] https://www.gesis.org/en/services/research/tools/thesaurus-for-the-social-sciences

[8] https://www.gesis.org/angebot/recherchieren/tools-zur-recherche/klassifikation-sozialwissenschaften (in German)

[9] http://www.openarchives.org

[10] https://jats.nlm.nih.gov

[11][spacy.io](spacy.io)

[12] http://methods.sagepub.com

[13] apart from those used in traditional NER systems like *Person*, *Location*, or *Organization* with abundant training data, as covered in the Stanford NER system(Finkel et al., 2005)

[14] https://nlp.stanford.edu/projects/project-ner.shtml

[15] https://coleridgeinitiative.org/richcontextcompetition with a total of 5,000 English documents

[16] https://acl-arc.comp.nus.edu.sg/

[17] Word embeddings are trained with a skip gram model using embedding size equal to 100, word window equal to 5, minimal occurrences of a word to be considered 10. Word embeddings are clustered using agglomerative clustering with a number of clusters set to 500,600,700 Ward linkage with euclidean distance is used to minimize the variance within the clusters.

[18] A glossary of statistical terms as provided in https://www.statistics.com/resources/glossary/ has been added as well.

Placeholder between chapters

Chapter Break

Placeholder between chapters

Chapter Break

abstract: |
The steadily increasing number of publications available to researchers makes it difficult to keep track of the state of the art. In particular, tracking the datasets used, topics addressed, experiments performed and results achieved by peers becomes increasingly tedious. Current academic search engines render a limited number of entries pertaining to this information. However, having this knowledge would be beneficial for researchers to become acquainted with all results and baselines relevant to the problems they aim to address. With our participation in the NYU Coleridge Initiative's Rich Context Competition, we aimed to provide approaches to automate the discovery of datasets, research fields and methods used in publications in the domain of Social Sciences. We trained an Entity Extraction model based on Conditional Random Fields and combined it with the results from a Simple Dataset Mention Search to detect datasets in an article. For the identification of Fields and Methods, we used word embeddings. In this paper, we present how our approaches performed, their limitations, some of the encountered challenges and our future agenda.

author:
- Rricha Jalota
- Nikit Srivastava
- Daniel Vollmers

- René Speck
- Michael Röder
- Ricardo Usbeck
- ‘Axel-Cyrille [Ngonga Ngomo]{}’

bibliography:
- ‘references.bib’

title: |
DICE @ Rich Context Competition 2018 – Combining Embeddings and
Conditional Random Fields for Research Dataset, Field and Method

```
Recognition and Linking
```

Literature Review

Previous works on information retrieval from scientific articles are mainly seen in the field of Bio-medical Sciences and Computer Science, with systems [@DBLP:journals/ploscb/WestergaardSTJB18] built using the MEDLINE[5] abstracts, full-text articles from PubMed Central[6] or ACL Anthology dataset[7]. The documents belonging to the above-mentioned datasets follow a similar format, and thus, several metadata and bibliographical extraction frameworks like CERMINE [@tkaczyk2014cermine] have been built on them. However, since articles belonging to the domain of Social Sciences do not follow a standard format, extracting key sections and metadata using already existing frameworks like GROBID [@lopez2009grobid], ScienceParse[8] or ParsCit [@councill2008parscit] did not seem as viable options, majorly because these systems were still under development and lacked certain desired features. Hence, building upon the approach of Westergaard et. al [@DBLP:journals/ploscb/WestergaardSTJB18], we built our own sections-extraction framework for dataset detection and research fields and methods identification.

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

The problem of dataset detection and methods and fields identification is not only somewhat 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

{width="\textwidth"}

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

Preprocessing {#preprocess}

The publications were provided in two formats: PDF and text. For Phase–1, we used the given text files, however during Phase–2, we came across many articles in the training files that had not been properly converted to text and contained mostly non-ASCII characters. To work with such articles, we relied on the open source tool `pdf2text` from `poppler suite`[9] to extract text from PDFs. The `pdf2text` command served as the first preprocessing step and was called as a subprocess from within a python script. It was used with `-nopgbrk` argument to generate the text files.

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

- Words split by hyphens were de-hyphenated
- Irrelevant data was removed (i.e., equations, tables, acknowledgment, references);
- Main sections (i.e., abstract, keywords, JEL-Classification, methodology/data, summary, conclusion) were identified and extracted;

- Noun phrases from these sections were extracted (using the python library, spaCy[10]).

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

If a section was not found in the article (because of no explicit mention), 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 them to prevent loss of any meaningful data. Table [tab:sections] shows the number of identified sections in validation data. For brevity, we have evaluated only four main sections: title, abstract, keywords and methodology/data, since these are the ones getting preferential treatment in methods and fields identification.

[C[4cm]{} C[3.5cm]{} C[4cm]{}]{} **Sections** & **No explicit mention**
& **Mentioned but not found**\
Title & 0 & 0\
Keywords & 13 & 2\
Abstract & 0 & 1\
Methodology/Data & 18 & 4 We used `pdfinfo` from zhe `poppler suite` to extract PDF metadata that
very often contained the keywords and subject of an article. This tool was helpful in those cases where the keywords were not found by the regular expression.\
In the end, the preprocessing module generated four text files for a publication: PDF-converted text, PDF-metadata, processed articles containing relevant data, and noun phrases from the relevant sections, respectively. These files were then passed on to the other two components of the pipeline, which have been discussed below.

Approach

Research Fields and Methods Identification

Vocabulary Generation and Model Preperation

1. **Research Methods Vocabulary**: In Phase–1 of the challenge, we used the given methods vocabulary. However, the feedback that we received from Phase–1 evaluation gave more emphasis to statistical methods used by the authors, references to the time scope, unit of observation, and regression equations rather than the means used to compile the data, i.e., surveys. Since the given methods vocabulary was not a complete representation of statistical methods and also consisted terms depicting surveys, in Phase–2, we decided to create our own Research Methods Vocabulary using Wikipedia and DBpedia.[11] We manually curated a list of all the relevant statistical methods from Wikipedia[12] 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. For examples, please refer Table [tab:vocab].

   [C[1.5cm]{} C[5cm]{} C[5cm]{}]{} **Label** & **Description** & **Noun Phrases from Description**\
   Political forecasting & Political forecasting aims at predicting the outcome of elections. & Political forecasting, the outcome, elections\
   Nested sampling algorithm & The nested sampling algorithm is a computational approach to the problem of comparing models in Bayesian statistics, developed in 2004 by physicist John Skilling. & algorithm, a computational approach, the problem, comparing models, Bayesian statistics, physicist John Skilling

2. **Research Fields Vocabulary**: For both the phases, we used the given research fields vocabulary and, just like the methods vocabulary, added noun phrases from the description of the labels to it. However, since our phase–1 model seemed to confuse fields with methods, for Phase–2, we additionally created a blacklist of terms that didn't contain any domain-specific information, such as; Mixed Methods, Meta Analysis, Narrative Analysis and the like.
3. **Word2Vec Model generation**: In this pre-processing step, we used the above-mentioned vocabulary files containing noun phrases to generate a vector model for both research fields and methods. The vector model was generated by using the labels and noun phrases from the description of the available research fields and methods to form a sum vector. The sum vector was basically the sum of all the vectors of the words present in a particular noun phrase. 3em [The pre-trained Word2Vec model `GoogleNews-vectors-negative300.bin` [@DBLP:journals/corr/abs–1301–3781] was used to extract the vectors of the individual words.]{}

4. **Research Method training results creation**: For research methods, we generated an intermediate result file for the publications present in the training data. It was generated using a `naïve finder algorithm` which, for each publication, selected the research method with the highest cosine similarity to any of its noun phrase's vectors. This file was later used to assign weights to research methods using Inverse Document Frequency.

Processing with Trained Models

- **Finding Research Fields and Methods:** To find the research fields and methods for a given list of publications, we performed the following steps: (At first, Step 1 was executed for all the publications, thereafter Step 2 and 3 were executed iteratively for each publication).
    1. **Naïve Research Method Finder run** - In this step, we executed the `naïve research method finding algorithm` (i.e. selected a research method based on the highest cosine similarity between vectors) against all the current publications and then merged the results with the existing result from the `research methods' preprocessing step`. The combined result was then used to generate IDF weight values for each `research method`, to compute the significance of recurring terms.
    2. **IDF-based Research Method Selection** - We re-ran the algorithm to find the closest research method to each noun phrase and then sorted the pairs based on their weighted cosine similarity. The weights were taken from the IDF values generated in the first step and the manual weights assigned (section-wise weighting). Here, the noun phrases that came from the methodology section and from the methods listed in JEL-classification (if present) were given a higher preference. The pair with the highest weighted cosine similarity was then chosen as the Research Method of the article.
    3. **Research Field Finder run** - In this step, we first found the closest research field from each noun phrase in the publication. Then we selected the Top N (= 10) pairs that had the highest weighted cosine similarity. Afterwards, the noun phrases that had a similarity score less than a given threshold (= 0.9) were filtered out. The end-result was then passed on to a post-processing algorithm.\
    For weighted cosine similarity, the weights were assigned manually based on the section of publication from which the noun phrases came. In general, noun phrases from title and keywords (if present) were given a higher preference than other sections, since usually these two sections hold the crux of an article. Note, if sections could not be discerned from an article, then noun phrases from the section, reduced_content (see section [preprocess]), were used to find both fields and methods.

4. **Research Field Selection** - The top-ranked term from the result of step 3, which was not present in the blacklist of irrelevant terms, was marked as the research field of the article.

Dataset Extraction

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

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

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

2. **Rasa-based Dataset Detection:** In our second approach, we trained an entity extraction model based on conditional random fields using Rasa NLU [@DBLP:journals/corr/abs–1712–05181]. For training the model we used the Spacy Tokenizer[13] for the preprocessing step. For Entity Recognition we used BILOU tagging and used 50 iterations to train the CRF. We used the Part of Speech tags, the case of the input tokens and the suffixes of the tokens as input features for the CRF model. We particularly tested two configurations for training the CRF-based NER model. In Phase–1, the 2500 labeled

publications from the training dataset were used for training the Rasa NLU[14] model. Later in Phase–2, when the Phase–1 holdout corpus was released, we combined its 5000 labeled publications with the previously given 2500 labeled publications and then retrained the model again with these 7500 labeled publications.\
Running the CRF-Model: The trained model was run against the preprocessed data to detect dataset citations and mentions. Only the entities that had a confidence score greater than a certain threshold value (= 0.72) were considered as dataset mentions. A dataset mention was considered as a citation only if it was found in the given Dataset Vocabulary (via string matching either with a dataset title or any of the terms in a dataset 'mention_list') and if it belonged to the research field of the article. To check if a dataset belonged to the field of research, we found the cosine similarity of the terms in the 'subjects' field of the Dataset Vocabulary with the keywords and the identified Research Field of the article.

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

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

Evaluation

We performed a quantitative evaluation for Dataset Extraction using the evaluation script provided by the competition organizers. This evaluation (see Table [tab:dataset]) was carried out against the validation data, wherein we compared four different configurations. As can be inferred from the table, there was only a slight increase in performance for the Rasa-based model, when the training samples were increased. However, combining it with the Simple Dataset Mention Search, increased the performance by *19.42%*. Interestingly, there was no improvement in performance in the combined approach even when the training samples for the Rasa-based model were increased. This might be because of the removal of frequently-occuring terms from the Rasa-generated output, based on the frequency distribution of dataset mentions as computed in the Simple Dataset Mention Search. M[2.2cm]{} | M[2.3cm]{} | M[2.2cm]{} M[2.2cm]{} M[2.2cm]{}]{} & &\

Metrics& **Rasa-based Approach** (2500) & **Rasa-based Approach**
(7500) & **Combined Approach** (2500) & **Combined Approach** (7500)\
Precision & 0.382 & 0.388 & **0.456** & **0.456**\
Recall & 0.26 & 0.26 & **0.31** & **0.31**\
F1 & 0.309 & 0.311 & **0.369** & **0.369** For Research Fields and Methods, we carried out a qualitative evaluation against 10 randomly selected articles from Phase–1 holdout corpus. Tables [tab:field] and [tab:method] depict a comparison between the predicted fields and methods in Phase–1 and Phase–2. In general, our models returned a more granular output in the second phase, solely because of the modifications we made in the vocabularies.

[C[1cm]{} C[4.5cm]{} C[3cm]{} C[3.5cm]{}]{} **pubid** & **Keywords** &
Phase–1 & **Phase–2**\
10328 & Cycling for transport, leisure and sport cyclists & Health
evaluation & **Public health and health promotion**\
7270 & Older adult drug users, harm reduction & Health Education &
Correctional health care\
6053 & Economic conditions - crime relationship, homicide & Homicide &
Gangs and crime [C[1cm]{} C[4.5cm]{} C[3cm]{} C[3.5cm]{}]{} **pubid** & **Keywords** &
Phase–1 & **Phase–2**\
10328 & Thematic content analysis & Thematic analysis & **Sidak
correction**\
7270 & Interviews conducted face to face, finding systematic patterns or
relationships among categories identified by reading the interview
transcript & Qualitative interviewing & **Sampling design**\
6053 & Autoregressive integrated moving average (ARIMA) time-series
model & Methodological pluralism & **Multivariate statistics**

Discussion

Throughout the course of this competition, we encountered several challenges and limitations in all the three stages of the pipeline. In the preprocessing step, the appropriate extraction of text from PDFs turned out to be rather challenging. This was especially due to the varied formats of the publications, which made the extraction of specific sections—that contained all data relevant to our work—demanding. As mentioned before, if there was no explicit mention of the key-terms like
`Abstract, Keywords, Introduction, Methodology/Data, Summary, Conclusion`
in the text, then the content was saved as ‘reduced_content’ after applying all other preprocessing steps and filtering out any irrelevant data.\
Our experiments suggest that the labeled publications we received for dataset detection were not uniform in the dataset mentions provided, which made it difficult to train an entity extraction model even with an increased number of training samples. Hence, there was only a slight

improvement in performance when the Rasa-model was trained with 7500 publications instead of 2500. This was also why we combined the Rasa-based approach with the Simple Dataset Mention Search, so that at least the datasets that were present in the vocabulary didn’t get missed.

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

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

1. For research methods, merging the given SAGE methods vocabulary with our manually curated vocabulary, could have resulted in methods that would have been both granular and statistical while still being relevant to the publications. Introducing a blacklist just as we did for research field identification, could also have been another workaround.
2. For both fields and methods identification, we could have also tried pre-trained embeddings from glove[15] and fastText[16].
3. As our entity-extraction approach for Dataset Detection suffered from a limitation of inconsistent labels (i.e. datasets mentioned in the form of abbreviation, full-name, collection procedure, and keywords) in training data, we could have extended the Simple Dataset Mention Search to a pattern-oriented search based on handcrafted rules derived from sentence structure and other heuristics.

Future Agenda

The data provided to us in the competition displayed a cornucopia of inconsistencies even after human processing. We hence propose that machine-aided methods for computing correct and complete structured representation of publications are of central importance for scientific research such as an Open Research Knowledge Graph [@DBLP:journals/corr/abs–1901–10816]. Previous works on never-ending learning have shown how humans and extraction algorithms can work together to achieve high-precision and high-recall knowledge extraction from unstructured sources. In our future work, we hence aim to populate a **scientific knowledge graphs** based on never-ending learning. The methodology we plan to develop will be domain-independent and rely on active learning to classify, extract, link and publish scientific research artifacts extracted from open-access papers. Inconsistency will be remedied by ontology-based checks learned from other publications such as SHACL constraints which can be manually or automatically added.[17] The resulting graphs will

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

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

Appendix

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

Chapter Break

author:
- |
Philips Kokoh Prasetyo, Amila Silva, Ee-Peng Lim, Palakorn Achananuparp\
Living Analytics Research Centre\
Singapore Management University\
[{pprasetyo,amilasilva,eplim,palakorna}@smu.edu.sg]{}

bibliography:
- ‘rcc–02.bib’

title: Simple Extraction for Social Science Publications

Abstract

With the vast number of datasets and literature collections available for research today, it is very difficult to keep track on the use of datasets and literature articles for scientific research and discovery. Many datasets and research work using them are left undiscovered and under-utilized due to the lack of available search tools to automatically find out who worked with the data, on

what research topics, using what research methods and generating what results. The Coleridge Rich Context Competition (RCC) therefore aims to build automated dataset discovery tools for analysing and searching social science research publications. In this paper, we describe our approach to solving the first phase of Coleridge Rich Context Competition.

Introduction

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

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

This manuscript describes summary of our submission for the first phase of RCC. We begin with related work in section [sec:relatedwork]. We present our analysis on RCC dataset in section [sec:data], describe our approach in section [sec:methods], and discuss our experiment results in section [sec:experiments]. Finally, we wrap up with conclusion and future work in section [sec:conclusion].

Related Work {#sec:relatedwork}

Extracting information from scientific text has been explored in the past [@Peng2004AccurateIE; @Nguyen2015ScholarlyDI; @Singh2016OCRAR]. One type of information extraction from scientific articles is extracting keyphrases and relation between them [@Augenstein2017SemEval2T]. @Luan2017ScientificIE propose semi-supervised sequence tagging approach to extract keyphrases. @Augenstein2017MultiTaskLO explore 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 [@Alves2018DeepRM], and citation metadata extraction [@Hetzner2008ASM; @Anzaroot2014LearningSL; @An2017CitationME]. @Nasar2018InformationEF write a survey on information extraction from scientific articles.

Recently, there are some work to explore dataset extraction from scientific text [@Boland2012IdentifyingRT; @Ghavimi2016ASA; @Ghavimi2016IdentifyingAI]. @Boland2012IdentifyingRT propose weakly supervised pattern induction to identify references in social science publications. @Ghavimi2016ASA [@Ghavimi2016IdentifyingAI] propose a semi automatic approach for detecting dataset references for social science texts. Dataset extraction is a challenging task because of the inconsistency and wide range of dataset mention styles in research publications [@Ghavimi2016IdentifyingAI].

Data Analysis {#sec:data}

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

Preprocessing. In order to reliably access important structures of paper publications, we parse all papers using AllenAI Science Parse[5] [@Ammar2018ConstructionOT]. AllenAI Science Parse reads PDF file, and returns title, authors, abstract, sections, and bibliography (references). Since this parser utilizes machine learning models to parse PDF file, the parsing results may not be 100% accurate. Furthermore, this parser is unable to parse scan copy of old publication. In the situation where we are unable to access parsed fields, we fall back to the given text files.

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

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

[tab:train_top_sections] Top 12 Sections Mentioning Datasets in Training Set

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

[tab:dev_top_sections] Top 12 Sections Mentioning Datasets in Development Set

Citation Analysis. We build citation network from training set. Each node in the network is a paper publication, and an edge between two node A and B is generated if a paper A cites paper B. Table [tab:network_stats] shows the statistics of the citation network.

Number of nodes	5,000
Number of edges	998
Network density	0.008%

[tab:network_stats] Statistics of Citation Network

Initially, we propose an approach utilizing citation network based on an intuition that datasets, research methods, and research fields are shared by: 1) same or similar issues, 2) same or similar context, 3) same or similar authors and communities, 4) same or similar metrics used in the publication. However, based on table [tab:network_stats], we learn that exploring rich context using paper-paper citation network is not viable at this stage because most papers listed in publications' bibliography are not available in the training set, and therefore, paper-paper citation network becomes very sparse with many unknown information. Due to this reason, we drop our idea on utilizing paper-paper citation graph at this stage. Nevertheless, we believe that bibliography contains important signals and information about datasets, and research fields.

Methods {#sec:methods}

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

Dataset Extraction {#ssec:dataset_extraction}

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

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

For dataset recognition, we use an implicit entity linking approach. We start with the Naive Bayes model, which can be regarded as a standard information retrieval baseline, and entity indicative weighting strategy is used to improve the model. In order to calculate the word distribution of each dataset, we represent each dataset using its title, dataset mentions (provided in the training set), and dataset relevant sentences, filtered from the relevant publications using the rule based approach proposed in @Ghavimi2016IdentifyingAI. All these text sections related to a particular dataset are considered as a single text chunk, and we calculate the word distribution as follows. Let $\mathbf{w}$ be the set of words in a dataset. In our problem setting, we assume the dataset prior probability $p(d)$ to be uniform. The probability of dataset d given $w \in \mathbf{w}$ is:

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

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

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

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

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

where E is the number of distinct datasets considered and $df(w)$ counts the number of datasets with at least one occurrence of w.

Then for a given unseen publication, we use same rule based approach [@Ghavimi2016IdentifyingAI] to filter a few relevant sentences, and datasets are ranked by $ln(p(d|w))$ to select the most suitable datasets. In order to select exact datasets related to particular publication, we select top 10 datasets ranked using above approach. And then the confidence probability related to the top 10 datasets are normalized and select the datasets with the normalized probability higher than a predefined threshold value. We return the entity indicative words as relevant dataset mentions.

Research Methods Identification {#ssec:research_method_identification}

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

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

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

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

Research Fields Identification {#ssec:research_field_identification}

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

To gain more understanding about the characteristic of each field, we crawl top search results from SAGE Knowledge[6]. From the search result snippets, we collect information such as title and abstract on various publications including case, major work, books, handbooks, and dictionary. We exclude video and encyclopedia. Due to sparseness of the SAGE Knowledge, we exclude all research fields with less than 10 search results. In the end, we have samples of 414 L3 research fields under 101 L2 research fields and 10 L1 research fields. This numbers cover 20.87% of 1,984 L3 research fields, and 67.79% of 149 L2 research fields in the list of SAGE research fields. We use this data to train research fields classifiers.

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

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

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

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

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

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

6. Return research fields having score $>= 0.4$.

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

Experiment Results {#sec:experiments}

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

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

—————————————————

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

—————————————————

[tab:dd_dev_result] Dataset Detection Results on Development Set

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

—————————————————-———

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

—————————————————-———

[tab:de_dev_result] Dataset Extraction Results on Development Set

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

[tab:de_test_result] Dataset Extraction Result on Test Set

Table [tab:de_test_result] shows dataset extraction performance on test set (phase 1). The significant drop from development set result suggests that the test set might have different distribution compare to the training and development set. It might also contain dataset citations that are never been seen in training set.

Research Methods Identification. We only consider Naive Bayes and Logistic Regression classifiers for research method identification because they naturally outputs probability score. We perform 5-fold cross validation to evaluate classification performance, and the result can be seen in table [tab:rmethods_5cv]. Logistic regression classifier outperforms Naive Bayes with 0.86 F1 score in classifying 133 research methods.

Classifier	F1
Naive Bayes	0.55
Logistic Regression	0.86

[tab:rmethods_5cv] F1 Score for Research Method Classification

Research Fields Identification. We perform 5-fold cross validation to evaluate our classifiers to classify L1, L2, and L3 research fields. Table [tab:rfields_pub_5cv] shows the results using n-gram features from paper title and abstract, whereas table [tab:rfields_rt_5cv] shows the results using n-gram features from title only. Naive Bayes tends to perform slightly better on L3 research fields where we have large number of research field labels. We decide to use SVM for research field identification on publication level because SVM is generally better than Naive Bayes. On the other hand, we decide to use Naive Bayes for research field identification on bibliography level because Naive Bayes prefer to have more accurate L2 and L3 research fields.

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

[tab:rfields_pub_5cv] F1 Score for Research Field Classification on Publication Level using Paper Title and Abstract

— — — — — — — — — — — — — — –

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

— — — — — — — — — — — — — — –

[tab:rfields_rt_5cv] F1 Score for Research Field Classification on Bibliography Level using Paper Title Only

Conclusion {#sec:conclusion}

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

— - –

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

Apart from F1 score on 5-fold cross validation, we have no good way to evaluate research method and research field identification without ground truth label. Our methods are unable to automatically extract and recognize new datasets, research methods, and fields. An extension to automatically handle such cases using advance Natural Language Processing (NLP) approach is a promising direction.

From this competition, we have learned that lacks of labelled training data is a huge challenge, and it directs us to other external resources (i.e., SAGE Knowledge) as proxy for our label. Another challenge is data sparsity. Although we see many paper listed in bibliography, lacks of access to these publication make us difficult to exploit citation network.

Unfortunately, our model did not advance to the second phase. We are interested in exploring more advanced information extraction methods on the RCC datasets, and we hope that the organizer will release the RCC datasets for future research. We thank the organizers for organizing a competition and workshop on this important, interesting, and challenging problem.

Chapter Break

Placeholder for Reseach Agenda and Next Steps chapter.—
abstract: |

Datasets are critical for scientific research, playing a role in replication, reproducibility, and efficiency. Researchers have recently shown that datasets are becoming more important for science to function properly, even serving as artifacts of study themselves. However, citing datasets is not a common or standard practice in spite of recent efforts by data repositories and funding agencies. This greatly affects our ability to track their usage and importance. A potential solution to this problem is to automatically extract dataset mentions from scientific articles. In this work, we propose to achieve such extraction by using a neural network based on a BiLSTM-CRF architecture. Our method achieves $F_1 = 0.883$ in social science articles released as part of the Rich Context Dataset. We discuss limitations of the current datasets and propose modifications to the model to be done in the future.

author:
- ‘Tong Zeng$^{\{1}$,2\}$andDanielAcuna^{\{1\}\$[}$1]’

bibliography:

- ‘rcc–06.bib’
title: |
Dataset mention extraction in scientific articles using a BiLSTM-CRF

`model`

Introduction

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

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

In this article, we describe a method for extracting dataset mentions based on a deep recurrent neural network. In particular, we used a Bidirectional Long short-term Memory (BiLSTM) sequence to sequence model

paired with a Conditional Random Field (CRF) inference mechanism. We tested our model on a novel dataset produced for the Rich Context Competition challenge. We achieve a relatively good performance of $F_1 = 0.883$. We discuss the current noise and duplication present in the dataset and limitations of our model.

The dataset

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

We adopt the CoNLL 2003 format [@tjong2003introduction] to annotate whether a token is a part of dataset mention. Concretely, we use we use B-DS denotes a token is the first token of a dataset mention, I-DS denote a token is inside of dataset mention, and O means a token is not a part of dataset mention. We then put each token and its corresponding labels in one line and use a empty line as separator between sentences. All the sentences was split by 70%, 15%, 15% as training set, validation set and testing set.

The Proposed Method

Overall view of the architecture

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

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

1. Input Layer: input the sequence of tokens to the network;
2. Embedding layer: mapping each token into fixed sized vector representation based on fasttext (200-dimensional vectors, @pennington2014glove)
3. One BiLSTM layer: make use of Bidirectional LSTM network to capture the high level representation of the whole token sequence input (200 dimensions per direction, totally 400 output units)
4. Dense layer: project the output of the previous layer to a low dimensional vector representation of the the distribution of labels.
5. CRF layer: find the most likely sequence of labels.

?{width="5cm"}

Word Embedding

The embedding is the first layer of our network and it is responsible for mapping the word from string into vectors of numbers as the input for other layers on top. For a given sentence S, we first convert it into a sequence consisting of n tokens, $S = \{c_1, c_2, \cdots, c_n, \}$. For each token $c_i\$welookuptheembeddingvector\$x_i$ from a word embedding matrix $M^{tkn} \in \mathbb{R}^{d|V|}$, where the d is the dimension of the embedding vector and the V is the Vocabulary of the tokens. In this paper, the matrix M^{tkn} is initialized by a pre-trained embedding, but will be updated by learning from our corpus.

LSTM

Recurrent neural network (RNN) is a powerful tool to capture features from sequential data, such as temporal series, and text. RNN could capture long-distance dependency in theory but it suffers from the gradient exploding/vanishing problems [@pascanu2013difficulty]. The Long short-term memory (LSTM) architecture was proposed by @hochreiter1997long and it is a variant of RNN which copes with the gradient problem. LSTM introduces several gates to control the proportion of information to forget from previous time steps and to pass to the next time step. Formally, LSTM could be described by the following equations:

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

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

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

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

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

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

where the σ is the sigmoid function, $\bigotimes$ denotes the dot product, b is the bias, W is the parameters, x_t is the input at time t, c_t is the LSTM cell state at time t and h_t is hidden state at time t. The i_t, f_t, o_t and g_t are named as input, forget, output and cell gates respectively, they control the informations to keep in its state and pass to next step.

LSTM get information from the previous steps, that is left context in our task. However, it is important to consider the information in the right context. A solution of this information need is bidirectional LSTM [@graves2013speech]. The idea of Bi-LSTM is using two LSTM layers and feed in each layer with sequence forwards and backwards separately, and then concatenate the hidden states of the two LSTM to modeling both the left and right contexts

$$h_t = [\overrightarrow{h_t}\text{\textbackslash varoplus}\overleftarrow{h_t}]$$

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

Results

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

For all of our results, we use F_1 as the measure of choice. This measure is the harmonic average of the precision and recall and it is the standard measure used in sequence labeling tasks. This metric varies from 0 to 1, and the unit is the highest possible value. Our method achieved a relatively high F_1 of 0.883 for detecting mentions, in line with previous studies.

We found significant limitations to the dataset, and we expect these limitations to affect the linkage step (not done in this article). While we are proposing a model for such step, we found that it would be challenging to do so given the quality of the annotations. Specifically, we found significant duplication of labels. The first issue is that mentions are nested (e.g. HRS, RAND HRS, HRS DATA, RAND HRS DATA are nested and linked to the same dataset). The second issue is that for the same mention text, several, different datasets were linked (e.g. the term CPS is linked to 57 datasets, the term NHANES is linked to 32 datasets). This adds noise to the linkage process. In fact, most of the mentions have ambiguous relationships to datasets. In particular, only 17,267 (16.99%) mentions are linked to one dataset, 15,292 (15.04%) mentions are listed to two datasets, and 12,624 (12.42%) are linked to three datasets. We found that there were some extreme cases, where for example there are several mentions linked to more than one hundred datasets. If these difficulties are not overcome, then the predictions from the linkage process will be noisy and therefore impossible to tell apart.

Conclusion

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

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

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

Acknowledgements {#acknowledgements .unnumbered}

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

---

1. The authors would like to thank Rafael Beier for helpful comments.We would like to thank Jannick Blaschke for providing the graphs in Section 2.2. The views expressed here do not necessarily reflect the opinion of Deutsche Bundesbank or the Eurosystem. ↩
2. Data producers in different departments across Bundesbank compile, e.g. microdata, indicators, or time series. ↩
3. In our model, we have called this knowledge user specific knowledge. Here the knowledge is in that sense specific, that it can be used to fulfil the task of Bundesbank in a better way ↩
4. For the future, ongoing effort is needed to support all four “corners of the circle” / “pillars of the level model”. The current competition strengthens the arrow from publication to knowledge and structures gained knowledge to improve data services. To support the data services pillar – for example -, digital RDC environments with facilitated access processes like a “data stewardship module” will in our view improve data access. ↩
5. Corresponding author: deacuna@syr.edu ↩
6. http://sk.sagepub.com/browse/ ↩

Placeholder between chapters

7. https://www.aclweb.org/anthology/ ↩
8. https://github.com/allenai/science-parse ↩
9. [poppler]https://manpages.debian.org/testing/poppler-utils ↩
0. https://github.com/explosion/spaCy ↩
1. https://wiki.dbpedia.org/services-resources/ontology ↩
2. https://en.wikipedia.org/wiki/Category:Statistical_methods ↩
3. https://spacy.io/api/tokenizer ↩
4. https://rasa.com/docs/nlu ↩
5. https://nlp.stanford.edu/projects/glove ↩
6. https://fasttext.cc/docs/en/crawl-vectors.html ↩
7. https://www.w3.org/TR/shacl/ ↩

Placeholder between

chapters

8. https://www.europeansocialsurvey.org/data/ ↩
9. The Australia National Data Service, https://www.ands.org.au/ ↩
0. There are additional collections at http://data.census.gov, http://gss.norc.org.]{.underline} http://electionstudies.org, http://psidonline.isr.umich.edu, and http://www.nlsinfo.org. ↩
1. https://www.earthcube.org/ ↩
2. https://pds.nasa.gov/ ↩
3. https://www.uniprot.org/ ↩
4. http://www.rcsb.org/ ↩
5. https://datadryad.org/ ↩
6. https://www.openarchives.org/pmh/ ↩
7. https://www.colectica.com/ ↩
8. https://gssdataexplorer.norc.org/ ↩
9. http://wiss-ki.eu ↩
0. http://www.dcc.ac.uk/ ↩
1. http://irods.org ↩
2. http://schema.org/dataset/datastreams ↩
3. There are two distinct ways the term micro-data is used. 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. ↩
4. https://mmisw.org/ ↩
5. https://www.re3data.org/ ↩
6. https://joinup.ec.europa.eu/release/dcat-ap/11 ↩
7. E.g., Online Analytical Processing (OLAP), Enterprise Data Warehouses (EDW), and Decision Support Systems (DSS). ↩
8. https://www.w3.org/TR/vocab-data-cube/ ↩
9. https://www.myexperiment.org/home ↩
0. https://www.crossref.org/ ↩
1. https://orcid.org/ ↩
2. https://duraspace.org/vivo/ ↩
3. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/ ↩

Placeholder between chapters