February 19, 2019

Stefan Bender, Hendrik Doll, Christian Hirsch

**Where’s Waldo: The search for data in empirical research**

Chapter: **Conceptual Issues**

Content

[1 Introduction 2](#_Toc1490571)

[1.1 Vision RDSC 2](#_Toc1490572)

[1.2 Roadmap: theory 2](#_Toc1490573)

[1.3 Roadmap: a first step 3](#_Toc1490574)

[1.4 RDSC interest in competition 3](#_Toc1490575)

[1.5 Competition fit into broad strategy Bundesbank 3](#_Toc1490576)

[1.6 Wrap-up – what to expect from paper 4](#_Toc1490577)

[2 Lessons learned from competition– a social scientists view 4](#_Toc1490578)

[2.1 What exactly is a dataset mention in a paper? 4](#_Toc1490579)

[2.1.1 Named vs. created datasets 4](#_Toc1490580)

[2.1.2 Domain specificity of dataset usage 5](#_Toc1490581)

[2.1.3 Datasets used for analysis vs. cited 5](#_Toc1490582)

[2.2 What are appropriate definitions for fields/methods? 5](#_Toc1490583)

[3 Conclusion 7](#_Toc1490584)

[3.1 Describe competition outcome 7](#_Toc1490585)

[3.2 General lesson learned #1: curated corpus publication is needed 7](#_Toc1490586)

[3.3 General lesson learned #2: DOIs for datasets would make life easier 7](#_Toc1490587)

[3.4 Way forward 7](#_Toc1490588)

## Introduction

## Vision RDSC

Deutsche Bundesbank as the German central bank offers free access to microdata for independent research on its premises through its research data and service centre (RDSC). Access restrictions arise, because data – on the one hand - is a public good, but – on the other hand– often is confidential. By granting access, the questions arise about the additional benefit of data provision for Bundesbank and society. So to speak, we are interested in the dataset impact factor.

By providing the best possible data information for researchers / *“providing a map of the dataset landscape*”, we enhance researcher knowledge of *“what is out there”* and thereby enable effective quantitative research through optimal data usage. This supports tackling societal challenges and research on policy relevant fields, which in turn leads to an effective data-driven policy decision making process.

## Roadmap: theory

We plan to provide added value to researchers about similar data sets to the ones they use, by building a dataset recommendation system. Given the multitude of available data sources within and outside the institution, we enable ameliorated research by providing researchers with the means to take a better informed decision about using all available data to the best extent.

We need to extract datasets used in research output and categorize datasets by “similarity” / joint usability to enable data-driven amazon.com style recommendations like *“people like you also used dataset x”.* Traditionally, dataset metadata (information on datasets) is provided from the producer perspective (data collection methods, universe, sampling design, regulatory requirements, etc.). Such data is relatively readily available, but of lesser use to fuel recommendations for three main reasons.

First, because data descriptions (of new data, which comes in ever-increasing frequencies) need ever-increasing manual curation. Second, because new datasets without annotated expert knowledge would not be considered close to existing datasets and recommended (thereby hindering fast spread of relevant new datasets). And third, because data might seem close from the producing side, but for a multitude of reasons cannot be used together in a meaningful way. For these reasons, we consider metadata from the user-side / output-side. As output from data, we understand research papers. This leads us to the need for new information: Which paper uses what data. For this, we extract mentioned data sets from research papers.

[Maybe add motivation about RDC use-case / metadata description from the user side and by the users?]

## Roadmap: a first step

From research papers as an unstructured data source, by text extraction, we get the relevant strings out of the paper referring to data set citations. Subsequently, entity resolution (match different data set names and descriptions relating to the same dataset) is applied. From the obtained knowledge on dataset usage, categorization of similar datasets by similar usage is made possible (usage by the same authors, usage of different datasets in the same research fields, analysis of different datasets by the same statistical methods).

For such a data-driven look at datasets from usage in research papers *(categorize datasets from the user side),* information on used datasets, fields and methods is necessary. Datasets are useful intrinsically, fields and methods to enable better useful recommendations. I.e. extracting datasets was the focus of the competition, fields and methods were the icing on the cake to allow better suggestions, based on predicting dataset joint usability if two datasets are analyzed in the same research field with the same methods.

## RDSC interest in competition

Against this background, the project at hand is a first, but major step in the direction of a user-centric dataset recommendation system and dataset impact evaluation. Before being able to measure the added value of Bundesbank data to research, it is necessary to find papers automatically, where Bundesbank data is used, and gather information about the used content and methods.

A machine learning competition is an appropriate format to show a proof of concept for obtaining the required data. Beforehand, the baseline knowledge for dataset usage in research publications was effectively zero (excluding islands of hand-annotated publications). We were interested in seeing, if world class data scientists and software engineers can show solutions to this absence of data, thereby laying the groundwork to allowing a recommendation system. Now, the information requirements for such an envisioned recommendation system are shown to be manageable and scalable. The competition is an important first step and proofs that data set extraction is possible in theory.

## Competition fit into broad strategy Bundesbank

In line with ongoing digitalization efforts, automated metadata extraction allows more readily the use of additional AI and machine learning algorithms. Specializing in solving the extraction task from unstructured data first, lifts the value unstructured, yet underexplored, data has. Transformation of data to easily analyzable and utilizable forms allows more digitally integrated processes (digitalization of RDSC / Bundesbank).

## Wrap-up – what to expect from paper

Our background is in social science and data science. From this perspective we think about optimizing data documentation, access and dissemination for research. We are no librarians by trade; we may be excused for sounding repetitive for hardened librarians. In the present paper, we present our findings on dataset citations, fields and mentions, what to look for, how to define, and how to categorize those.

## Lessons learned from competition– a social scientists view

## What exactly is a dataset mention in a paper?

### Named vs. created datasets

Datasets in social science can be categorized into two broad categories. First, there are named datasets, i.e. well defined and publicized datasets (e.g. Compustat). Usually, they are cited as short strings, often containing institution name or name of commercial data vendor and they are usually well-defined in scope and time. Often they have an often used abbreviation (e.g. MMSR). While data collection is usually subject specific, data can be used across multiple papers and potentially research domains

Pitfalls are multiple subsamples / waves of same datasets, which makes it difficult to exactly identify the same used data. [Add stuff to DOI and cittations from Paper with Gesis.] Issues of definitions are if different time periods or subsamples refer to the same data set entity. Usage of a digital object identifier (DOI) [add reference here] offers a solution, but thus far is not universal for research data sets. With a DOI (identifying the exact data version) data sets are identified, solving the problem at hand, and quantitative research is reproducible.

Without a DOI, even if the extraction algorithms yield exact identification of named datasets, reproducibility [add references from Vilhuber presentation] is not always ensured, because state of data and knowledge among other issues is not well-defined. Named data is thus not first best (for this DOI or additionally identify and precisely define wave, used filters, coding steps or identify final used dataset), but second best for our purposes.

As a different category of data sets, there are created datasets. Those are often described in one ore multiple paragraphs together with data collection and sampling methods. There is usually no specific string, which allows unique identification for the human reader. Data collection is usually paper specific, and existing datasets are not easily searchable. This makes data collection potentially redundant and data spread not optimal. There is no well-accepted reference repository, containing information, which survey datasets exist. Specific created datasets are harder to use for follow-up research, and reproducibility is given only if publishers hold available the data together with the paper.

### Domain specificity of dataset usage

Across social science domains, fractions vary how widespread named datasets are used as opposed to created datasets. Also, the number of datasets per empirical paper varies across domains and dependent on named vs. created. In fields with widespread use of multiple datasets at once (linked data), the added value of recommending additional useful data is expected to be higher than in fields that create specific study-specific data every time. [also the opposite argument could be constructed].

### Datasets used for analysis vs. cited

Dataset mentions in papers consist of two types. First, datasets used for the empirical analysis and second, cited datasets in the literature review or references. In some cases it is not clear where to draw the line for cited datasets. E.g. some report statistics based on datasets (i.e. they mention statistics from cited papers e.g. “Author x uses Compustat to…”. Sometimes differences 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 make up our minds ahead of time, what information to look for, used or cited datasets. Which information should be extracted depends on the purposes of the use-case in mind. Note that in a first best setting, if information were available for the universe of datasets used for analysis in papers and all paper citations would be available, dataset citations would be redundant.

While literature citations are mostly standardized within research domains and are relatively straightforward to extract (hence publication networks / publications maps exist), information on used datasets in papers remains incomplete (even after the competition). Because of this, for the competition, we asked for used and cited datasets. It is important to note, that 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 dataset citations, one obtains a “dataset map”, thus the “impact of datasets”. Through real dataset usage, one obtains the relevant information for our purposes, namely information relating to dataset similarity and joint usage possibilities from the user perspective.

Another lesson learned is to also include theoretical literature, essays, etc. in the corpus of publications to present to the algorithms. It is helpful if algorithms 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 this distinction.

## What are appropriate definitions for fields/methods?

For the competition, on purpose, we did not provide any thesauri. The rationale behind this, was to see the unhindered creativity of teams, which available information sources they would use or not use (reference datasets, Wikipedia, other repositories, thesauri, statistical clustering techniques, etc..).

Thesauri limit the catalogue of potentially identifiable fields and methods, thus prohibiting new methods and fields to be identified in fast-changing modern research areas. Also thesauri might disturb algorithm performance, since algorithm might be forced to categorize topics and fields to older / less exact categories than necessary. On the other hand, the reasons for using thesauri are well-known [add reference?] and include easy clustering of similar fields and methods and a manageable category set of predictions.

[Add here Discussion archive.org / discussion using Wikipedia articles as a reference]

Concerning our definition of methods for the purpose of the competition two questions arise. First, what type of methods should be extracted from the paper? There are statistical methods, sampling methods, qualitative methods, etc., where do we draw the line? Second, if the answer is statistical method, which statistical method in a paper should be extracted? E.g. in an empirical paper, is only the main causal analysis relevant or also methods for data wrangling, sampling, robustness checks, descriptive statistics, etc?

Although there are different types of datasets and types of citations, there is a fair understanding of what a dataset is (structured / unstructured, named / created, …). However, there is no clear-cut definition of what fields and methods should be. With these degrees of freedom in mind, we have to define those with our specific use-case in mind. For useful new recommendations to be provided to researchers, it is necessary to define methods as methods that allow a merge of datasets / joint usability. We learned to define categories based on the task to solve.

From a dataset joint usage perspective, important methods are those that distinguish, whether two datasets can be linked or analysed together. With this in mind, we propose to 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 regression equations. If two papers then use different datasets in the same field using the same methods, there is a relatively high likelihood that those datasets might be linked or used together to create new insights

Another lesson learned is that we generally face a fine line between too broad predictions (safe, but uninformative) and too narrow predictions (narrow, but potentially wrong). A potential way out is backward induction here – we can present differently aggregated predictions for fields and methods to users and get feedback from them (let users rank usability – *“Was this helpful to you?”*)

## Conclusion

## Describe competition outcome

Competition was great first step, proofs extraction is possible in theory, and was important for us to learn. [Add reference to motivation above, added value etc.]

## General lesson learned #1: curated corpus publication is needed

Better data is needed. No curated corpus is readily available in decent quality that is ready to use. This is a potential explanation, why dataset usage in publications is not systematically captured and analysed yet. If no such dataset is available, try to manually annotate papers. In the course of the evaluation, we realized open methodological questions and clarified many of our needs for precise definitions, what to look for.

## General lesson learned #2: DOIs for datasets would make life easier

In a perfect world, there would be universal DOI usage for unique identification of datasets and standardized dataset citations (much like paper citations). Since DOIs only gain widespread nowadays, our task is a 1:n mapping of publications to datasets without unique identifiers. For scientific papers many journals already provide DOIs, thereby allowing easy identification of papers and paper citation networks (for discussion papers, essays, policy briefs, newspaper articles, etc less so). There are ongoing efforts by journals to have all used data published for reproducibility reasons, but thus far there are no incentives for researchers to provide unique identification of datasets used in papers.

There is no way to enforce data providers to give DOIs for datasets. However, this article makes the case for more universal usage of the DOI for datasets. In the meantime, the present article shows a way forward to learn from the current literature that does not include widespread data set DOIs [numbers to back-up this claim?] and still analytically use what information is there. However, replication / reproducibility is not always ensured by finding names of used datasets.

## Way forward

[pick-up Motivation aufgreifen and show “the great picture“. Reference introduction motivation]

[What do competition results mean for overall FDSZ/Bundesbank goal]

###################################################################################

ENDE DES TEXTES

###################################################################################