March 15, 2019

Stefan Bender, Hendrik Doll, Christian Hirsch

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

Chapter: **Who’s Waldo: Conceptual issues when characterizing data in empirical research**

**Abstract**

Empirical economic and social science research uses microdata for analyses to connect theory to societal problems. At present, discussion centers on the impact of this research – for example as evidence based policy evaluation or analysis –, while little is known about the added value of the used (micro) data. Research data is often considered a public good and much data is generated by publically funded researchers or is available from official statistics (like administrative data). In turn, microdata should help to create new insights of challenges for society.

We present conceptual lessons learned from a machine learning competition held to automate the discovery of datasets, research methods and fields in research publications. Put differently, before you ask where Waldo is, you have to identify, who Waldo is. The obtained information can be used to build up a user-centric dataset recommendation and dataset impact evaluation system. Our main conclusion is that the research community is able to build such a system. There is a strong need to establish a data sharing culture, encouraged by giving credit where credit is due through dataset citations.

[1. Introduction 2](#_Toc3560261)

[2. Insights from a research data centre perspective 4](#_Toc3560262)

[2.1 Background 4](#_Toc3560263)

[2.2 A model of the empirical knowledge generating process 4](#_Toc3560264)

[2.3 Added value of structured user-specific knowledge 7](#_Toc3560265)

[3. Lessons learned from competition 8](#_Toc3560266)

[3.1 Dataset mentions 8](#_Toc3560267)

[Lesson #1: datasets fall into two broad categories 9](#_Toc3560268)

[Lesson #2: Fractions of dataset category are domain specific 9](#_Toc3560269)

[Lesson #3: Unique identification of datasets remains an issue 10](#_Toc3560270)

[Lesson #4: Datasets mentions could indicate used for analysis vs. cited 11](#_Toc3560271)

[3.2 Fields and Methods 12](#_Toc3560272)

[Lesson #5: Think before you act: define fields and methods 12](#_Toc3560273)

[3.3 Next steps 13](#_Toc3560274)

[4. Conclusions 14](#_Toc3560275)

[5. References 15](#_Toc3560276)

# Introduction

Policy makers increasingly recognize that informed decision-making – for example – in the field of monetary policy requires microdata on financial institutions and markets to uncover interdependencies between entities and document global developments. Making microdata available entails costs which are related to the need to safeguarding statistical confidentiality and making accompanying metadata available.

At Deutsche Bundesbank (the German central bank), the Research Data and Service Centre (RDSC) offers access to microdata free of charge for independent research. As one of the largest data producers in Germany, Bundesbank offers access to its high quality micro data for research purposes. Since access to microdata is subject to legal and data protection requirements, Bundesbank set up the RDSC, which provides researchers and analysts with access to selected micro data in the context of independent scientific research projects.

To strengthen effective quantitative research through optimal microdata usage, the RDSC has engaged in a series of projects that are targeted at i) quantifying the marginal benefit of microdata provision for society in general and Deutsche Bundesbank in particular, and ii) enhancing user experience. The competition at hand is a first, but major step towards producing a framework for a user-centric dataset recommendation system and dataset impact evaluation. As such, the competition is intended to showcase the feasibility of such a comprehensive approach.

One specific project currently pursued by the RDSC is the development of a microdata map, which will serve as an important input for a data-driven recommendation system for researchers. Traditionally, metadata on microdata is provided from a producer’s perspective (e.g. data collection methods, universe, sampling design, regulatory requirements, etc.). The envisioned, user-centric, data driven recommendation system, is novel in terms of providing information about how other researchers have actually used microdata in previous research.

While an important part of the dataset discovery process and for informing researchers, traditionally compiled metadata is insufficient to recommend microdata to researchers in the finance domain for two reasons. First, the discovery of newly compiled microdata may be slow because these datasets typically do not have annotated expert knowledge, when they first become available for research. Second, while microdata might seem similar from the data producer side, researchers may venture to link these datasets for empirical research.

Therefore, a challenge in building a data-driven recommendation system is to make sure that recommended datasets 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.

Empirical research papers are a natural source of information about dataset use. However, as any researcher or librarian may confirm, hand-curating a large corpus of research papers is labor-intensive and error-prone. Therefore, being able to automatically retrieve the necessary information from research papers lays the groundwork for any future dataset recommendation system. The competition is an important first step and proofs that data set extraction is shown to be manageable and scalable.

Extracting dataset citations from publications is a fairly difficult task because of the variety of dataset citation formats and the absence of training data (for a recent overview 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.

As Boland and Mathiak (2015) describe dataset extraction as a twofold task, finding dataset citation string and following entity resolution (match the string to the correct entity/ DOI). Concerning entity resolution, they report the difficulty of broad survey dataset citations that ignore data variability (such as years, versions, questionnaire variants, etc.), motivating a dataset taxonomy. Named dataset citations are often underspecified allowing identification of the survey but not of the precise dataset (which of multiple subsamples, 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 variety of fields for our purposes. To tackle this, we continue with a larger hand-curated annotated corpus.

Metadata schemas for datasets are available, such as 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.

Improving dataset citation is high on the scientific agenda in recent years. This notably includes promoting widespread usage of persistent and unique dataset identifiers. As available datasets are 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.

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.

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. The results can become a blueprint for future projects, where information is less structured.

# Insights from a research data centre perspective

## Background

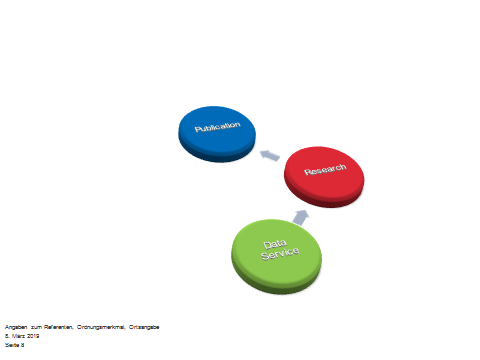
In response to the increased demand for microdata, in 2013 the Bundesbank set up the Integrated Microdata - based Information and Analysis System (IMIDIAS). One component of IMIDIAS is the recently established Research Data and Service Centre (RDSC). The RDSC applies a standardised procedure to generate high-quality datasets that cover a large part of data requests for research purposes. Thereby the RDSC grants internal and external researchers access − subject to clear conditions − to selected microdata provided by Deutsche Bundesbank, 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 anonymized datasets on banks, securities, investment funds, enterprises and households, all of which can be accessed at dedicated researcher workstations or for most of the Bundesbank’s surveys – as for the Panel on Household Finances (PHF) study – the RDSC offers so called scientific use files.

In addition, the RDSC provides advice 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 recognized standards and was accredited as a research data centre (RDC) by the German Data Forum(“Rat für Sozial- und Wirtschaftsdaten”).

## A model of the empirical knowledge generating process

In this section, we first describe a model of the empirical knowledge generating process in a RDC and relate this model to the status quo. We then proceed by describing, how and more importantly where the outcome of the competition fits into this model. In a last step, we provide examples of how the outcome of the competition may help to improve the status quo. While we use the RDC research process as an example, we believe that the empirical knowledge generating process we present in this section is a fairly general model which is applicable to other institutional settings as well.



Currently, in the RDSC – after a successful approval process – access to microdata is given. In figure 1, this data input to empirical research is labelled data services. Data services consist of the microdata itself and of information on data (metadata) from the data producing units[[1]](#footnote-1) in Bundesbank and the RDSC. With the help of these units, the RDSC documents, enriches and annotates data and facilitates access to microdata for research. The resulting product from data services thus goes beyond raw data. Refined, cleaned and linked data is provided, along with counsel if needed, as input to empirical research.

Researchers as users of micro data and data services (for example metadata), produce research outcomes. These outcomes – after data confidentiality clearance – sometimes take the form of publications. Publications contain knowledge accumulated by researchers about data usage over time (experience), e.g., knowledge about dataset particularities. Unstructured knowledge refers to information not written down, not shared outside of the current project, not machine-reusable or not understandable by the interest public. Examples of such knowledge acquired by researchers include:

1. How data is used (e.g. additional data cleaning, variable transformation, combining datasets, using additional information)
2. What purposes data is used for (e.g. topics, methodology, research area)
3. What kinds of analyses or techniques have been tried and are used ultimately
4. What information is most valuable to get the results.

The challenge is to have a feedback loop from research and publications – as the users and outcome of microdata – back to data services to improve the data and/or the metadata (see figure 2). At the moment, this feedback loop is not present in a systematic way. The aim of the competition was to identify appropriate procedures to close this gap, by 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 ultimately enable better data services which in turn will make research outcomes more efficient and better in general.



Important to note is that, while publications provide knowledge about research outcomes, results are presented to the interested public in a form optimized for human consumability as unstructured text. “User Specific Knowledge” in figure 2 emphasizes knowledge that is findable and transformed for the different purposes. This requires unstructured text to be put into structured chunks of information for automatic classification and analyses. Only thereby can the information from the publication best be built upon. In order to get to such useful knowledge, we have to find (extract) and model (structure) information from research papers.

Summarizing, in this model the empirical knowledge generating process starts with data services. Research is conducted using such services. Researchers produce publications as the research outcome, from which information can be structured into analysable form. Data services can then be automatically and manually improved through feedback from the user-side. We arrive at the starting point (data services) but on a higher level. In turn better data services allow better research. By partly automatizing the feedback-loop between research, publications, knowledge, and data services, the knowledge generating circle can spin faster and be augmented more quickly. Thinking of the data-driven knowledge generating process, we consider a circle model that increases in levels.

## Added value of structured user-specific knowledge

Being able to access structured user-specific knowledge as described in the last section, 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 researchers. Similarly, data producers benefit from feedback on their data, allowing them to improve data quality.

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

Going further, the usability index can be expanded into a measure, how well new datasets fit each other. For this proposals are predicted based on dataset usage in the same field, using the same methods or by additional metadata similarity (hence without necessarily needing joint dataset usage in the past). This can be a valuable accelerator to effectively distribute new datasets in the research community. Both indices can be implemented using information from the competition only, however we propose extensions, based on other information such as current metadata, that may enhance value to users.

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

In our setting with researchers and datasets, the universe of users/ researchers is decently large, but per user, we observe a limited amount of “dataset consumption”. 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 needs based on his past direct “dataset shopping” behaviour.

As a hypothesis, we suspect a simpler underlying behavioural model of “data shopping”, because publishing with one dataset is not a casual purchase. It implies real commitment and also successful publishing, which relates to being content with the purchase (less cognitive dissonance). Thus, we would need less data points per person than online platforms, 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.

The RDSC also has a responsibility towards its principals i.e. society as a public institution. Granting data access free of charge for researchers should be backed by empirical benefits of such data provision. Thereby societal investment in free data access can be justified. Societal benefits of data access can be measured in knowledge[[2]](#footnote-2) created by specific datasets. Knowing which datasets underlie new research findings (which datasets are used in which publication and which findings are enabled by the specific dataset), enables evaluating the added value of a dataset for Bundesbank or society.

Measuring societal benefits through data is not obvious at first glance. Judging the added value of dataset provision first requires identifying which empirical result from a publication can be attributed to which dataset. One can argue that added value of providing micro data is the marginal benefit compared to the second-best data, if comparable commercial databases exist. 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. In its most extreme case, new datasets can allow entirely new research questions to be answered.

This section detailed two examples of value generated by the competition for the RDSC. First, systems can be optimized in a user-centric way, by obtaining refined products (e.g. improved researcher recommendations and data documentation). Second, with dataset usage data, the case for societal investment in free data access can be empirically fortified. Positive externalities (research as a public good) would currently suggest a less then societally optimal provision of processes and institutions that generate research data and offer related services. In the next section, we proceed by describing what we learned in the competition on the road to data-driven systems.

# Lessons learned from competition

## Dataset mentions

This section presents lessons 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, education, and others), we quickly realized that our dataset image had an econocentric bias. In social science, we learned, datasets can be categorized into two broad categories for the purposes of extraction. First, there are named datasets, i.e. well defined, usually large-scale and publicized datasets (e.g. Compustat).

Generally, named dataset mentions are short strings in the publications, have commonly used abbreviations (e.g. MMSR), and often containing institution name or name of commercial data vendor. Sometimes (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 papers and research domains.

The second dataset category is what we call created datasets. By created dataset we understand datasets usually collected or built by authors of a publication for the purpose of analysing one specific research question. Often, created data comes in the form of small-scale surveys, (structured) interviews, or randomized controlled trials, RCTs. Such data normally does not have a trademark name, but instead 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 commonly used abbreviation. Data collection is usually paper specific, and the universe of existing datasets are not easily searchable. This makes it hard for text mining algorithms to correctly extract strings referring to dataset entities. Specific created datasets are harder to use for follow-up research, and reproducibility is given only if publishers provide data together with the paper. Therefore, the lack of unique identification and search terms renders data collection potentially redundant and dataset spread not optimal.

### Lesson #2: Fractions of dataset category are domain specific

Throughout the competition 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 identification techniques and differing potentials for conducting RCTs, the predominantly used data sources naturally vary. This has important repercussions for designing a competition, since algorithm performance and later recommendation system performance varies with the input corpus and the application field.

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

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

### Lesson #3: Unique identification of datasets remains an issue

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

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

With a DOI (identifying the exact time frame, sampling universe, data version, wave, aggregations, state of knowledge, etc.), datasets are identified and quantitative research using confidential microdata is reproducible. To make lives easier, DOIs also 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 straightforward to extract (hence publication networks / publications maps exist), information on used datasets in papers remains incomplete (even after the competition). Because of this, for the competition, we asked for used and cited datasets. It is important to note, that extracted dataset citations are always incomplete, since some authors report aggregate statistics from a different paper, but not the data behind (*“Smith et al show…”*).

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

As training data for the algorithms it is important to include theoretical literature, essays, etc. in the corpus of publications. 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 competition, on purpose. The rationale behind this was to see the unhindered creativity of teams, which available information sources they would use or not use (e.g. reference datasets, Wikipedia, archive.org, other repositories, thesauri, statistical clustering techniques, etc.). On the other hand, thesauri limit the catalogue of potentially identifiable fields and methods, thus prohibiting new methods and fields to be identified in fast-changing modern research areas. Also thesauri might disturb algorithm performance, since algorithm might be forced to categorize topics and fields to older or less exact categories than necessary.

However, using thesauri does have well-known advantages, as any librarian will confirm. These advantages include easy clustering of similar fields and methods and a manageable category set of predictions. For field predictions, we generally face a fine line between too broad predictions (safe, but 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 to users and get feedback from them (let users rank usability – *“Was this helpful to you?”*).

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

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

## Next steps

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

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

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

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

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

We introduce a circular model of the knowledge generating process, which increases in levels. From data services, research is conducted, publications are published and user-specific knowledge is generated. Having such knowledge on dataset usage, data services in turn can be improved. 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.[[3]](#footnote-3)

Automatic processing of generated knowledge in publications becomes increasingly available with modern text analysis tools. Extracting such information is important, because timely and optimal usage of gained results increases the speed, by which findings can be incorporated into data services and thereby next-level research is enabled in turn. To further improve automatic processing, minimum standards for dataset taxonomy are needed. Harmonized metadata schemas for data sets – like the INEXDA metadata schema (Bender et al. 2018) for central banks and statistical offices (compliant with and building upon DataCite) – offer such an approach.

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

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

# References

* + Ball, A., and M. Duke (2011): How to cite datasets and link to publications. Digital Curation Centre.
  + Bender, S., Hausstein, B., & C. Hirsch (2018). An Introduction to INEXDA’s Metadata Schema. Technical Report 2018-02, Deutsche Bundesbank, Research Data and Service Centre.
  + Boland, K., Ritze D., Eckert, K., & B. Mathiak (2012): Identifying references to datasets in publications. 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
  + 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.
  + 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.

1. Data producers in different departments across Bundesbank compile – for example – microdata, indicators, or time series. [↑](#footnote-ref-1)
2. 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 [↑](#footnote-ref-2)
3. 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. [↑](#footnote-ref-3)