A Semi-Automatic Approach for Detecting Dataset References in Social Science Texts

Behnam GHAVIMI [a,b,1], Philipp MAYR [a,2], Christoph LANGE [b,c,3], Sahar VAHDATI [b,4] and Sören AUER [b,c,5]

[a] *GESIS – Leibniz Institute for the Social Sciences*
[b] *Enterprise Information Systems (EIS), University of Bonn*
[c] *Fraunhofer Institute for Intelligent Analysis and Information Systems IAIS*

Abstract. Today, full texts of scientific articles are often stored in different locations than the used datasets. Dataset registries aim at a closer integration by making datasets citable but authors typically refer to datasets using inconsistent abbreviations and heterogeneous metadata (e.g. title, publication year). It is thus hard to reproduce research results, to access datasets for further analysis, and to determine the impact of a dataset. Manually detecting references to datasets in scientific papers is time-consuming and requires expert knowledge in the underlying research domain. We propose and evaluate a semi-automatic three-step approach for finding explicit references to datasets in social sciences papers.

We first extract pre-defined special features from dataset titles in the da|ra registry, then detect references to datasets using the extracted features, and finally match the references found with corresponding dataset titles. The approach does not require a corpus of articles (avoiding the cold start problem) and performs well on a test corpus. We achieved an F-measure of 0.84 for detecting references in full-texts and an F-measure of 0.83 for finding correct matches of detected references in the da|ra dataset registry.

Keywords. Information extraction, Link discovery, Data linking, Research data, Social Sciences, Scientific papers, Data registry

1. Introduction

Today, many articles reference research datasets which they are based on. For example, an article might present the results of a statistical analysis performed on a dataset comprising employment data in the EU. However, dataset references are usually not explicitly exposed in digital libraries. In most cases the articles do not provide explicit links that give readers direct access to the referenced datasets.

[1]E-mail: behnam.ghavimi@gesis.org
[2]E-mail: philipp.mayr@gesis.org
[3]E-mail: langec@cs.uni-bonn.de
[4]E-mail: vahdati@cs.uni-bonn.de
[5]E-mail: auer@cs.uni-bonn.de

Explicit links from scientific publications to the underlying datasets and vice versa can be useful in multiple use cases, including:

- reviewers aiming to reproduce the evaluation that authors performed on a dataset,
- other researchers desiring to perform further analysis on a dataset that was used in a paper,
- decision makers seeking to determine the impact of a given dataset or to identify the most used datasets in a given community.

Currently, the majority of published papers lack such direct links to datasets. While there exist registries that make datasets citable, e.g., by assigning a digital object identifier (DOI) to them, they are usually not integrated with authoring tools. Therefore, in practice, authors typically cite datasets by *mentioning* them, e.g., using combinations of title, abbreviation and year of publication (see, e.g., Mathiak and Boland [1]).

Manually detecting references to datasets in papers is time consuming and requires expert knowledge of the paper's domain. Detecting dataset references automatically is challenging due to the wide variety of styles of dataset citations in full texts even within one research community, and the variety of places in which datasets can be referenced in papers (illustrated by Figure 1). So it is a difficult task to create a training set manually

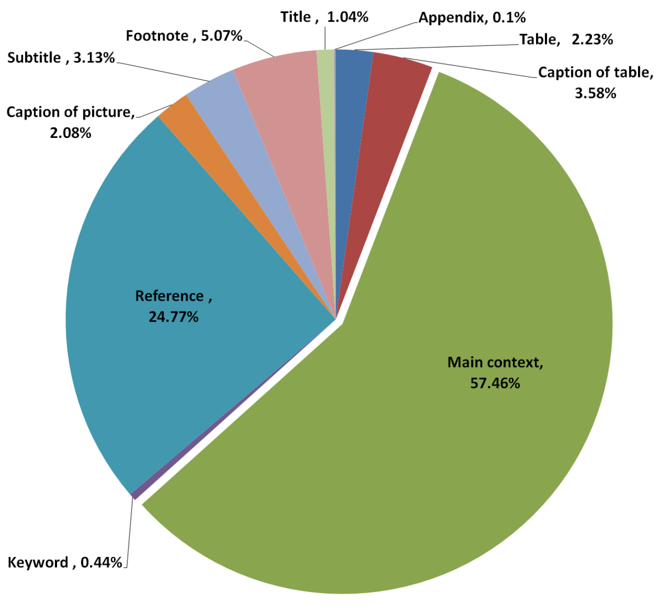

Figure 1. The distribution of more than 500 dataset references in 15 random papers from the mda journal (see Section 4.1)

for solving this issue and this variance even makes rule-based approaches difficult, as it is hard to cover all cases. We therefore introduce a semi-automatic approach that parses full texts and finds exact matches with a high precision without requiring a training set. The remainder of this section states more precisely the problem addressed by our research.

Section 2 introduces the preliminaries of techniques and metrics that we apply. Section 3 reviews existing literature related to this problem. All data used by our approach is explained in section 4, while section 5 introduces the proposed solution. The evaluation is then presented in section 6, and section 7 concludes with an outlook to future work.

1.1. Problem Statement

Whereas a lot of effort has been spent on information extraction in general [2], few attempts have focused on the specific use case of dataset extraction (see, e.g., [3]). When referring to the same dataset, different authors often use different names or keywords. Although there are proposed standards for dataset citation in full texts, researchers still ignore or neglect such standards (see, e.g., [4]). Since the structure of scientific documents does not always follow standard, and since datasets are rarely being linked to in a consistent ways, simple keyword or name extraction approaches do not solve the problem [5]. Table 1 shows concrete examples of different reference styles from five papers from the mda journal (see Section 4.1) that have cited different versions of a study (ALLBUS/GGSS = Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/German General Social Survey).

SV@BG:effort on information extraction and dataset extraction is not comparable, one is a research field, the other is a use-case, rephrase it.
CL: I made the first step: not yet perfect, but better.

CL@BG: say more about these papers. At the very least what journal they are from. Even better would be a footnote with citations of all papers. Journals value precision and completeness.
BG: I mentioned the journal in the text and I added DOIs to the table

	Paper DOI	Citation style
Paper A	10.12758/mda.2013.014	ALLBUS (2010)
Paper B	10.12758/mda.2013.004	GESIS – Leibniz-Institute for the Social Sciences: ALLBUS 2010 – German General Social Survey. GESIS, Cologne, Germany, ZA4610 Data File version 1.0.0. (2011–05–30), doi:10.4232/1.10445.
Paper C	10.12758/mda.2013.012	ALLBUS (Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften)
Paper D	10.12758/mda.2014.007	(e.g., in the German General Social Survey, ALLBUS; see Wasmer, Scholz, Blohm, Walter and Jutz, 2012)
Paper E	10.12758/mda.2013.019	Die Einstellungen zu Geschlechterrollen wurden mit Hilfe von Items aus den ALLBUS – Wellen 1994 und 2008 operationalisiert.

Table 1. Citation styles for a study in five different papers.

The challenge that our research aims at is to turn each dataset reference detected in a paper into an explicit link, for example, using the DOI of the dataset's entry in a dataset registry. Dataset registries usually provide further metadata about datasets, which can facilitate the detection of such links, such as the dataset's creators, publication date, description, and temporal coverage. In our case, references to datasets should be linked to items in the da|ra registry, which covers datasets from the social sciences and economics.

1.2. Contribution

This paper makes the following contributions:

- a quantitative analysis of typical naming patterns used in the titles of social sciences datasets,
- a semi-automatic approach for finding references to datasets in social sciences papers with two alternative interactive disambiguation workflows, and
- an evaluation of the implementation of our approach on a testbed of journal articles.

2. Preliminaries: similarity, ranking and evaluation metrics

Our work and the related work of other researchers employ certain terminology and standard metrics for ranking the results of a search query (here: a text in a paper that refers to a dataset) over a corpus of documents (here: titles of datasets), and for evaluating the accuracy of information retrieval algorithms. The following four subsections introduce the terminology and definitions of concepts used in this paper.

2.1. Dataset (Terminology)

Dataset is an ambiguous term; different authors have suggested a great variety of meanings for it [6]. Therefore, Renear et al. introduced a general notion of the term based on the definitions in the technical and scientific literature. They define a dataset in terms of its four basic characteristics *grouping*, *content*, *relatedness*, and *purpose*.

In their definition, a dataset is considered as a *group* of data. *Set*, *collection*, *aggregation*, and *atomic unit* are some cases of this feature type. For instance, a dataset may have *set* semantics (e.g. "Set of RDF triples") [7] so then it follows the mathematical definition of set, or it may have *collection* semantics so the deletion and addition of data do not have any effect on the dataset's identity. The *content* feature describes the data in a dataset. For example, *observation* describes propositional content, while *value* refers to measurable content.

CL@BG: please phrase more clearly; I find "not lose or accept" hard to understand.
BG@CL: I read the part in the referenced paper and I replaced it with a new sentence. Please check the new one.

A dataset is a group of data, which are *related* to each other. The *relatedness* feature thus clarifies the inter-relation of data in a dataset. *Syntactic* and *semantic* are some examples of this feature. Finally, the *purpose* feature refers to the idea of the scientific activity which the dataset is created for.

CL@BG: Is this classification only about research data, or could it be more general? In the latter case, how about rephrasing to "use case" and/or "application domain"?
BG:We can say Scientific "activity" (same term used in the reference).

2.2. Weighting terms in documents using tf-idf

The *bag of words* model represents a text as a set of terms without considering their order. Based on this model, documents and the query can be displayed in different ways, such as binary vectors, count vectors, and weight vectors.

CL@BG: Does your algorithm employ each of these vectors? If not, it would make sense to shorten this section to just those vector representations that you use.
BG: The extra parts were removed

In a weight matrix, each row represents a term and each column represents the vector of a document or query. Each cell in a weight matrix represents the weight of a term in a document. Tf-idf is one way of computing the weights.

Term frequency (tf) measures the number of occurrences of a given term (t) in a given document (d) or query text [8]. Tf is calculated by the following formula.

$$w_{t,d} = \begin{cases} 1 + \log_{10} tf_{t,d}, & \text{if } tf_{t,d} > 0 \\ 0, & \text{if } tf_{t,d} = 0 \end{cases}$$

The reason for using a logarithm in the formula is that a high number of occurrences of a term does not make the document linearly more relevant. The total weight for a document is calculated by summing the weights of all terms in the document. These terms should appear in both q and d. It is zero if none of the query terms exist in the document.

$$tf\_score = \sum_{t \in q \cap d} w_{t,d}$$

df_t is the number of documents in the corpus that contain t, so the more a term is repeated in the corpus, the term less informative it becomes. This reason leads to a new measure, which is idf *(inverse document frequency)*. Idf is effective for queries that have more than one term. The following formula defines idf, where N is the number of documents in the corpus.

$$idf = \log_{10}(N/df_t)$$

Tf-idf is defined as the product of *tf* and *idf*.

$$tf\text{-}idf(q,d) = \sum_{t \in q \cap d} tf.idf_{t,d}$$

When ranking documents that contain a term being searched, tf-idf returns high scores for documents for which the given term is *characteristic*, i.e. documents that have many occurrences of the term, while the term has a low occurrence rate in *all* documents of the corpus. In other words, tf-idf assigns a weight to each word in a document, giving high weights to keywords and low weights to frequent words such as stop words.

2.3. The cosine similarity metric

A Boolean search enables users to find patterns: if a document matches the pattern, the result is “true”; otherwise the result is “false”. In other words, documents either do or do not satisfy a query expression. A *ranked* retrieval model is more sophisticated, in that it returns a ranked list of documents in a corpus by considering a query.

Similarity measures such as Matching, Dice, Overlap Coefficient, and Jaccard are some examples of approaches for ranking a list of documents within a query (cf. Manning and Schütze [9]). Matching Coefficient finds the numbers of terms that occur in both the query and document vectors. It calculates the cardinality of the intersection of each document and the query.

$$MatchingCoefficient = |d \cap q|$$

Dice Coefficient, Overlap Coefficient, and Jaccard try to normalize the Matching Coefficient in different ways:

- $DiceCoefficient = \frac{2|d \cap q|}{|d|+|q|}$
- $OverlapCoefficient = \frac{|d \cap q|}{\min(|d|,|q|)}$

- $JaccardCoefficient = \frac{|d \cap q|}{|d \cup q|}$

Each of them has specific shortcomings; for example, Jaccard neither applies optimal normalization on the length of documents nor considers term frequency in a document and the corpus of documents. A document can be converted into a weight vector (see Section 2.2), which looks like $d = (w_1, \ldots, w_n)$, and tf-idf is one way of computing the weight w_i of terms.

Search results for a multi-word query in a corpus of documents can be ranked by the similarity of each document with the query. Given a query vector q and a document vector d, their *cosine similarity* is defined as the cosine of the angle θ between the two vectors [8, 9], i.e.

$$\cos(\overrightarrow{q}, \overrightarrow{d}) = \cos\theta = \frac{\overrightarrow{q} \cdot \overrightarrow{d}}{\|\overrightarrow{q}\| \, \|\overrightarrow{d}\|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

It normalizes vectors by converting them to their unit vector, thus making documents of different lengths comparable. Since Euclidean distance is not effective for vectors of different lengths, it ranks documents by angle instead of distance. Between 0 and 180 degrees, the cosine function decreases monotonically, and therefore larger angles mean less similarity. Combining tf-idf and cosine similarity yields a ranked list of documents. In practice, it may furthermore be necessary to define a cut-off threshold in order to distinguish documents that are considered to match the query from those that do not [10].

2.4. Precision and recall of a classifier

We aim at implementing a binary classifier that tells us whether or not a certain dataset has been referenced by a paper. The algorithm should find references of datasets in a paper and then detect a exact match for each detected reference in a text. These matches are to be selected from titles of datasets in a given dataset registry.

Evaluation metrics such as *precision and recall* determine the reliability of binary classifiers. For computing a unidimensional ranking of these two dimensions, one typically uses the F-measure, which is defined as the harmonic mean of precision and recall. These three metrics are defined as follows [11].

- $Precision = \frac{\text{\#True positives}}{\text{\#True positives} + \text{\#False positives}}$
- $Recall = \frac{\text{\#True positives}}{\text{\#True positives} + \text{\#False negatives}}$
- $F\text{-}measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$

If an algorithm returns few wrong predictions, it will lead to a high precision. An algorithm should predict most of the relevant results to achieve a high recall.

3. Related work

While only a few works address the specific task of extracting dataset references from scientific publications, a lot of research has been done on its general foundations including metadata extraction and string similarity algorithms. Related work can be divided into three main groups covered by the following subsections.

3.1. Methods based on the "bag of words" model

One group of methods is based on the "bag of words" model using algorithms such as tf-idf to adjust weights for terms in a representation of texts as vectors (cf. the introduction in Section 2.2). Lee and Kim proposed an unsupervised keyword extraction method by using a tf-idf model with some heuristics [2008]. Our approach uses similarity measures for finding a perfect match for each dataset reference in a paper by comparing titles of datasets in a registry to sentences in papers. Similarity measures such as Matching, Dice, Jaccard and Cosine can be applied to a vector representation of a text easily (cf. Manning and Schütze [9]). The accuracy of algorithms based on such similarity measures can be improved by making them semantics-aware, e.g., representing a set of synonyms as a single vector space dimension.

3.2. Corpus and Web based methods

Corpus and Web based methods often use information about the co-occurrence of two texts in documents, and are used for measuring texts' semantic similarity. Turney introduced a simple unsupervised learning algorithm for detecting synonyms [2001], which searches queries through an online search engine and analyses the results. The quality of the algorithm depends on the number of search results returned.

Singhal and Srivastava proposed an approach to extract dataset names from articles [2013]. They employed the normalized google distance algorithm, which estimates both the probability of two terms existing separately in a document, as well as of their co-occurrence.

$$NGD(x,y) = \frac{\max(\log f(x), f(y)) - \log f(x,y)}{\log M - \min(\log f(x), \log f(y))}$$

In this formula, M is the number of all web pages searched. $F(x)$ means the number of returned pages for x as a query term and $f(x,y)$ represents the number of pages for the intersection of x and y. They used two academic search engines – Google Scholar and Microsoft Academic Search – instead of a local corpus.

Schaefer et al. proposed the Normalized Relevance Distance (NRD) [2014]. This metric measures the semantic relatedness of terms. NRD is based on the co-occurrence of terms in documents, and extends NGD by using relevance weights of terms. The quality of these methods depends on the size of the corpus used.

Sahami and Heilman suggest a similarity function based on query expansion [2006]. Their algorithm determines the degree of semantic similarity between two phrases. Each of these phrases is searched by an online search engine and then expanded by using returned documents. Afterwards, the new phrases are used for computing similarity.

The problem that we aim to solve involves the two subtasks of 1. identifying dataset references in a paper, and then 2. finding at least one correct match for each of these identified references. Literature citation mining is the process of determining the number of citations that a specific paper receives. It constructs a literature citation network, which can be used for detecting the impact factor of a paper [17]. Citation mining can usually be handled by three subtasks. First, literature references should be extracted from the bibliography section of a document, and afterwards, metadata extraction should be

CL@BG: The combination of some author/year citations and some numeric citations, like here, looks strange to me. Please switch everything to *one* citation style: the one preferred by the journal.

applied on the references extracted in the first phase. Finally, each reference should be linked to the cited paper by using the metadata extracted in the second step [17].

Dataset and literature citation mining from documents cannot generally be compared in the detecting phase, since dataset mining needs to be applied to the entire paper, but literature mining must only consider the bibliography of the paper, unless the paper employs a "full" citation style. Unlike the detection phase, they can mostly use the same strategy for the matching phase.

CL@all: I added this new point; please check.

Afzal et al. proposed a rule-based citation mining technique [17]. Their approach detects literature references from each document and then extracts citation metadata from each of them, such as title, authors, and venue. Based on the venue, it then extracts all related titles from the DBLP computer science bibliography, which contains more than three million papers. Finally, it tries to link the title of each extracted literature reference and titles found in DBLP. Our approach tries to match dataset references in a paper to the titles of datasets registered in the da|ra registry.

CL@BG: consistently use either "registry" or "repository".
BG: I have applied the change for dara. I used "registry"

3.3. *Machine learning methods*

Many different machine learning approaches have been employed for extracting metadata, and in a few cases also for detecting dataset references. For example, Zhang et al. [2006] and Han et al. [2003] proposed keyword extraction methods based on support vector machines (SVM).

Kaur and Gupta conducted a survey on several effective keyword extraction techniques, such as selection based on informative features, position weight, and conditional random field (CRF) algorithms [2010]. Extracting keywords from a paper can be considered as a labeling task. CRF classifiers can assign labels to sequences of input, and, for instance, define which parts in a paper can be assumed to be keywords [21].

Cui and Chen proposed an approach using Hidden Markov Models (HMM) to extract metadata from texts [2010]. HMM is a language-independent and trainable algorithm [23]. Marinai described a method for extracting metadata from documents by using a neural classifier [2009]. Kern et al. proposed an algorithm that uses a maximum entropy classifier for extracting metadata from scientific papers [25]. Lu et al. used the feature-based Llama classifier for detecting dataset references in documents [2012]. Since there are many different styles for referencing datasets, large training sets are necessary for these approaches.

CL@BG: For this and possibly other references try to get the metadata as correct and complete as possible. Journals are picky about this. Here, find out the authors' full names. Also note that this one is actually an *@article*; use the fields *journal*, *volume* and *number*.

Boland et al. proposed a pattern induction method for extracting dataset references from documents in order to overcome the necessity of such a large training set [2012]. Their algorithm starts with either the name of a dataset or with an abbreviation of this name, and then drives patterns of all phrases that contain that name or abbreviation in papers. The patterns are applied to papers in order to extract more dataset names and abbreviations. This process repeats with new abbreviations and names until no more datasets can be detected in papers. It derives patterns of phrases that contain dataset references iteratively by using a bootstrapping approach.

4. Data sources

This section describes the three types of data sources that we use. We use full-text articles from the mda journal to evaluate the performance of our dataset linking approach, and metadata of datasets in the da|ra dataset registry to identify datasets. Finally, we use metadata of the papers registered in the SSOAR repository[6] for exporting the dataset links suggested for a paper in the JSON exchange format.

4.1. Papers from mda journal

Methods, data, analyses (mda[7]) is an open access journal that addresses questions important to quantitative methods, with a special emphasis on survey methodology. It has published research on all aspects of science of surveys, be it on data collection, measurement, or data analysis and statistics.For our research we used a random sample of full-text articles from mda as our test corpus.

4.2. The da|ra dataset registry

4.2.1. da|ra overview

Our proposed approach aims at social sciences datasets since it uses registered datasets in da|ra registry[8]. The registry offers a DOI registration service for datasets in the social sciences and economics. Different institutions have collected research data in the social sciences and made them available. Although the accessibility of such datasets for further analyses and other reuse is important, information on where to find and how to access them is often missing in papers.

da|ra makes data referable by assigning a DOI to each dataset. At the present time, da|ra holds 432,312 records such as datasets, texts, collections, videos, and interactive resources, 32,858 of which are datasets. For each dataset, da|ra provides metadata including title, author, language, and publisher. This metadata is exposed to harvesters employing a freely accessible API using OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting[9]).

CL@BG/PM: Once more this is IMHO a bit too much advertising **da|ra**. Can't we simply write "**da|ra** assigns a DOI to each dataset"?
BG: improved

CL@BG: Add a footnote with the date on which these figures were valid.

4.2.2. Analysis of dataset titles in da|ra

We analyzed the titles of all datasets in da|ra after harvesting them by using the da|ra API. The analysis shows that about one third of the titles follow a special pattern, which makes them easier to be detected in the text of a paper. We have identified three such special patterns. First, there are titles that contain *abbreviations*, which are often used to refer to the datasets. Consider, for example, the full title "Programme for the International Assessment of Adult Competencies (PIAAC), Cyprus", which contains the abbreviation "PIAAC". Secondly, there are *filenames*, as in the example "Southern Education and Racial Discrimination, 1880–1910: Virginia: VIRGPT2.DAT", where "VIRGPT2.DAT" is the name of the dataset file. Finally, there are *phrases* that explicitly denote the existence of dataset references in a text, such as "Exit Poll" or "Probation Survey". "Czech Exit Poll 1996" is an example of such a dataset title.

We assume these three categories as special characteristics in the titles. Abbreviations and special phrases can be found in about 17 and 19 percent of the da|ra dataset

[6] `http://www.ssoar.info`
[7] `http://www.gesis.org/en/publications/journals/mda/`
[8] `http://www.da-ra.de`
[9] `http://da-ra.de/oaip/`

titles respectively. The intersection of these two groups only accounts for 1.49 percent. Filenames occur in less than one percent of the titles. The proposed approach in this paper uses only the first and the last category, since the filename category only covers a small amount of titles.

4.3. The Social Science Open Access Repository (SSOAR)

The SSOAR repository provides full-text access to documents. It covers scholarly contributions related to different social science fields such as social psychology, communication sciences, and historical social research. Today, approx. 38,000 full-texts are available. This repository provides some metadata such as abstract and keywords for each paper in both German and English. It is assumed as a secondary publisher which publishes pre-prints, post-prints, and original publishers' versions of scholarly papers but it also lets authors publish their work for the first time.

Furthermore, the metadata of the papers inside the repository can be harvested easily. SSOAR assigns a URN (Uniform Resource Name) as a persistent identifier (PID) to each full-text to establish a stable link to the paper. If the full-text is the pre-print or post-print version of a published work, the repository uses the DOI of the paper.

5. A semi-automatic approach for finding dataset references

We have created a semi-automatic approach for finding references to datasets registered in da|ra in a given full text. Our approach is divided into four main steps. The first step is related to generating special features dictionaries from datasets' titles. The second step deals with identifying and matching datasets' references in a paper, and the third step focuses on improving the results of the second step. In the final step a user can export the results.

It took a semi-automatic approach since the first and last steps of our algorithm require human interaction to improve the accuracy of the result. In the first step, the user should review two generated lists of abbreviations and special phrases. In the final step, the user should make the final decision regarding references suggested by our approach.

The main differences between our approach and the other related works are that ours do not need a huge corpus of papers or a large training set. Our approach is straightforward and able to prepare results for a paper in few minutes or even seconds depending on the number of datasets' references in the paper that we want to analyze.

5.1. Step 1: Preparing the dictionary

The preparation of a *dictionary* of abbreviations and special phrases is the first step. *Abbreviations* are initially obtained by applying some algorithms and rules to the dataset titles harvested from da|ra. The dataset titles are preprocessed automatically before the abbreviations are extracted. Titles fully in capital letters are removed, the remaining titles are split based on ":", and then only the first parts are kept (in the case of including any colon mark). The extraction of abbreviations from dataset titles follows specific steps:

1. The titles are tokenized (by using nltk – a Python package for natural language processing).

2. The tokens that are not completely in lowercase (not including the first letter) – not only a combination of digits and punctuation marks, not Roman numerals, and do not start with a digit are added to a new list (e.g. "SFB580-B2", "A*CENSUS", "L.A.FANS", "aDvANCE" and "GBF/DIME").
3. The titles are split based on '-' and '(', and then single tokens before such delimiters are added to the list (e.g "euandi" in "euandi (Experteninterviews) - Reduzierte Version").
4. The items on the list of abbreviations should only contain the punctuation marks '.', '-', '/', '*' and '&'. (e.g. "NHM&E").
5. The items that contain '/' or '-' and are also partially in lowercase are removed from the list (first letter of each part is not included) (e.g "Allbus/GGSS" is removed).
6. Words in German and English, as well as country names, are removed from the list. Words, fully or partially in capital letters will not be pruned by dictionary (first letter is not included).

The titles fully in capital letters are converted into lowercase and tokenized. Afterwards, the dictionary prunes them, and then their tokens without definition are added to the list. These algorithms and rules correctly detect, for example, "DAWN" in "Drug Abuse Warning Network (DAWN), 2008". However, it sometimes detects abbreviations that are not references to datasets, such as "NYPD" in "New York Police Department (NYPD) Stop, Question, and Frisk Database, 2006". As their identification is hard to automate, we assigned this task to a human expert. The expert reviews the list and then makes a false positive list – such false positives will be removed from the dictionary automatically. The ratio of false positives is one out every three items on the list of abbreviations extracted automatically. This means that approximately 66 percent of the abbreviations are derived correctly, and the rest of the titles need very little effort in order to be pruned from the list.

The preparation of the dictionary of special phrases also needs human interaction. A list of terms that refer to datasets such as "Study" or "Survey" has been generated manually; this list contains about 30 items. Afterwards, phrases containing these terms were derived by some algorithms and rules from the titles of actual datasets in da|ra. Three types of phrases are considered here, the first of which are tokens that include an item in the dictionary such as "Singularisierungsstudie" where the phrase contains "studie". The second is a category of phrases that includes "Survey of" or "Study of" as a sub phrase as well as one more token that is not a stop word, such as "Survey of Hunting". The last one is phrases that contain two tokens, where one of them is an item in the dictionary such as "Poll", and the second token should not be a stop word such as "Freedom Poll".

A human expert has finally verified the phrase list, and false positives are added to the related list. In the phrase list, there are few false positives, and most can be detected while processing papers. This means our approach will improve over time. Both dictionaries – abbreviations and phrases – can be generated on the first harvest of dataset titles from da|ra, and they can be required to update with every subsequent harvest. The delta update feature is not implemented in the paper, but the idea makes our approach even faster.

5.2. *Step 2: Detecting dataset references and ranking matching datasets*

Next, the characteristic features (abbreviations or phrases) of dataset titles are detected in the full text of a given paper. A paper is split into sentences, and each of these features is searched for in each sentence. Any detection of the special features in a text means a dataset reference in the text exists. A sentence is split into smaller pieces if a feature repeats inside the sentence more than once, since such a sentence may contain references to different versions of a dataset. Any phrase identified in this step might correspond to more than one dataset title.

For example, "ALLBUS"[10] is an abbreviation for a famous social science dataset, of which more than 150 versions are registered in da|ra. These versions have different titles and, for instance, the titles differ from year of study such as "German General Social Survey – ALLBUS 1998", "German General Social Survey – ALLBUS 2010", and "German General Social Survey (ALLBUS) – Cumulation 1980–2012". In another example, two titles that both contain the "PIAAC" abbreviation are "Programme for the International Assessment of Adult Competencies (PIAAC), Cyprus" and "Programme for the International Assessment of Adult Competencies (PIAAC), Germany", i.e., two datasets that differ in their geographic coverage. The last example is the two versions of "EVS" dataset, "EVS – European Values Study 1999 – Italy" and "European Values Study 2008: Azerbaijan (EVS 2008) ", which differ in both their year of study and geographic coverage.

We solve the problem of identifying the most likely datasets referenced in the paper by ranking their titles with a combination of tf-idf and cosine similarity. In this ranking algorithm, we apply the definitions of Section 2, where the query is a candidate dataset reference found in the paper and the documents are the titles of all datasets in da|ra. It means that our approach tries to identify the most similar dataset title in the da|ra repository with a sentence that contains any of the special features where the sentence belongs to the analyzed paper.

5.3. *Step 3: Heuristics to Improve Ranking in Step 2*

For each reference detected in the full text of a paper we compute tf-idf over the full text and the list of dataset titles in da|ra, which contain a specific characteristic feature (abbreviation or phrase) detected in the reference. As it leads to many false positives based on our observation, comparing datasets' titles with a sentence in a paper, and, afterwards, ranking titles based on their score was not useful. Therefore we solved the problems by involving special features.

Our approach considers only the list of titles that contain the special feature detected in the reference, since they are related titles and the rest of the titles in the registry are irrelevant. We limit our options in order to improve the accuracy of our approach. We decided to use the list of titles and whole sentences of the paper, and not only the reference sentence, since this consideration enables us to have a bigger corpus of documents and to reach a better weight for each word. The utilization of titles that contain the special features reduces the weight score of the feature and raises the weight scores of other terms in the reference sentence. It therefore has a positive impact on accuracy.

[10] Allgemeine Bevölkerungsumfrage der Sozialwissenschaften = German General Social Survey

While a corpus of papers is typically huge, the size of all da|ra dataset titles and the size of the full text of an average paper are less than 4 MB each. Given this limited corpus size, our algorithm may detect some false keywords in a query, thus adversely affecting the result. For instance, Figure 2 illustrates a toy example of this problem.

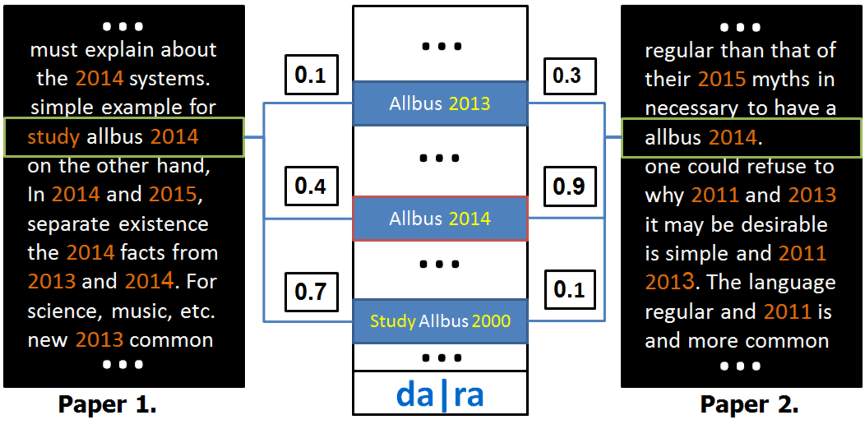

Figure 2. A toy example of cosine similarity, where tf-idf is computed over phrases in two papers.

In paper 1, “2014” repeats many times, whereas the word “study” occurs only once, which means the tf-idf assigns a high weight to “study” and a low weight to “2014”. When the query string is “study allbus 2014”, cosine similarity gives a higher rank to “Study Allbus 2000” than “Allbus 2014”.

To address this problem in a better way, our implementation employs some heuristics. This includes an algorithm that improves dataset rankings based on matching years in the candidate strings in both the paper and the datasets’ titles. In the example, these heuristics improve the ranking of the “Allbus 2014” dataset when analyzing paper 1.

Figure 3 shows an overview of our approach. Two steps labeled with “M” means they need human interactions. “M1” is about the preparation of lists of special features and “M2” is about making final decisions between candidates suggested by our approach.

5.4. Step 4: Exposing the Results to the User, and Interactive Disambiguation

The application of our approach supports two workflows through which an expert user can choose the best matches for the datasets cited by a paper from a set of candidates identified automatically. The sizes of these sets have been chosen according to the observations we made during the evaluation of the automated step, as explained in Section 6.

One workflow works per reference: for each reference, five titles of candidate datasets are suggested to the user. While this workflow best supports the user in getting every reference right, it can be time consuming. Each paper in our corpus contains 45 dataset references on average, but these *references* only belong to an average number of 3 distinct *datasets*.

The second alternative workflow takes advantage of this observation. It works per characteristic feature and suggests 6 titles of candidate datasets to the user for each feature (which may be common to multiple individual references in the paper).

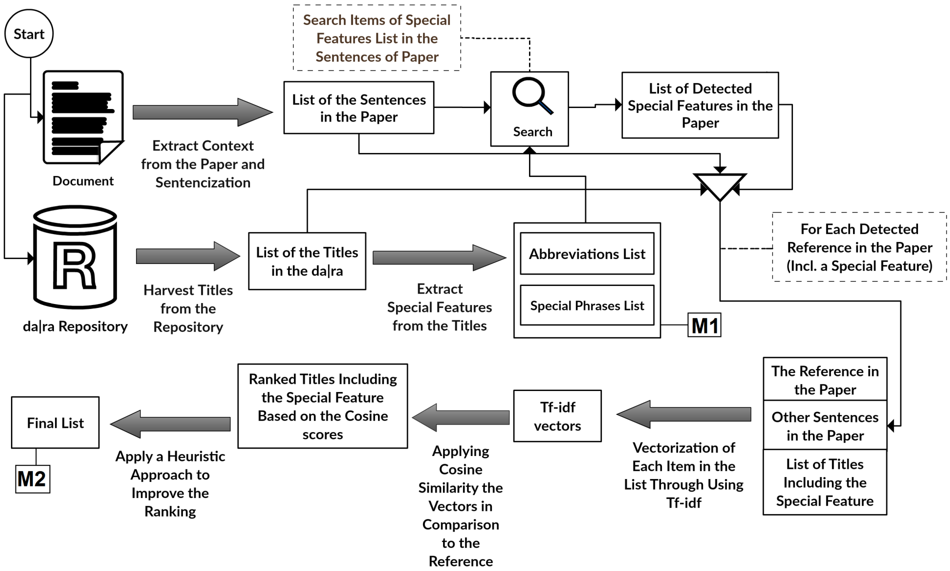

Figure 3. An overview of the approach.

Finally, an RDF graph which contains information about links identified between papers and datasets will be exported as an output. To enable even further analysis of the links identified between papers and datasets, we export an RDF graph containing all candidate datasets identified in the latter workflow for each paper. For each candidate dataset, we represent the essential metadata of the dataset in RDF: DOI and title.

6. Evaluation

The calculation of evaluation metrics such as precision, recall, and F-measure require a ground truth. We therefore selected a test corpus of 15 random papers from the 2013 and 2014 issues of the mda journal (see Section 4.1) – 6 in English and 9 in German. This test corpus includes 25 datasets without considering different versions of each dataset (see more information about the test corpus in table 2).

A trained assessor from the InFoliS II project at GESIS reviewed all papers one by one and identified all references to datasets (see second row in table 2). Afterwards, the assessor attempted to discover at least one correct match in da|ra for each detected reference, resulting in a lists of correct datasets per paper. These lists were used as a gold standard to compare with the results of our algorithm in order to examine differences and similarities.

,

6.1. Evaluation Process and Description

We decided to divide our evaluation into two steps. The first step focuses on identifying dataset references in papers. Here, accuracy depends on the quality of the generated

	Max.	Min.	Avg.
# of datasets in a paper	7	1	4
# of references to datasets	147	1	12

Table 2. Test corpus

dictionaries of abbreviations and special phrases (the accuracy metrics used in the paper are explained in 2.4).

Our algorithm searches these characteristic features (as explained in 5.2) in the full texts; detection of any of these features may lead to the detection of a dataset reference (see row "Detection" in table 3). In this phase, if a characteristic feature is identified both in a paper and in the gold standard, it will be labeled as a true positive. If the feature is in the gold standard but not in our output, it will be labeled as a false negative, or as a false positive in the opposite case.

The second step of the evaluation is about the accuracy of matching detected references in papers with datasets titles in the da|ra registry. This evaluation phase considers only true positives from the previous step. The lists of suggested matches for an item, both from the gold standard and from our output, are compared in this step. Since a dataset may occur on its own or be integrated together with other studies, an item can have more than one true match (e.g. Allbus 2010 in ALLBUScompact 1980–2012). In this step, an item will be labeled as a false negative if none of the suggestions for the item in the gold standard appears in the output of our algorithm. The number of false positives and false negatives are equal in this step, since a missing corresponding match means the possession of false positives. True positives, false positives, and false negatives are counted and then used to compute precision and recall.

The third row in table 3 refers to the accuracy of two phases of the algorithm as one unit in order to find how well it works generally, and does not consider one specific section (i.e. identification or matching). In order to satisfy this purpose, we repeated the second phase of evaluation, but this time included all data from the first step and not only the true positives. If an item is identified as false positive in the first section of evaluation, it is labeled as such in the evaluation as well.

6.2. Evaluation Results

The algorithm gains high precision in both the detection and matching phases, which means it has a much smaller number of wrong predictions. It also covers the majority of relevant data, which leads to high recall. The results of evaluations that we calculated are shown in table 3.

Our observations in the second evaluation step confirm the choices of set size in the interactive disambiguation workflows. In the per-reference matching workflow (as mentioned in 5.4), a ranked list of dataset titles is generated for each of the 45 dataset references (on average in our corpus) in a paper by employing a combination of cosine similarity and tf-idf.

Our observation shows that the correct match among da|ra dataset titles for each reference detected is in the top 5 items of the ranked list generated by combining cosine

Phase of Evaluation	Precision	Recall	F-measure
Detection	0.91	0.77	0.84
Matching	0.83	0.83	0.83
Detection+Matching	0.76	0.64	0.7

Table 3. Results of the Evaluation

similarity and tf-idf for that reference. Therefore, we adjusted our implementation to only keep the top 5 items of each candidate list for further analysis, such as an expert user's interactive selection of *the* right dataset for a reference.

The per-feature matching workflow (as mentioned in 5.4) categorizes references by characteristic features. For example, in a paper that contains exactly three detected characteristic features – "ALLBUS", "PIAAC", and "exit poll" – each dataset reference relates to one of these three features. If we obtain for each such reference the list of top 5 matches as in the per-reference workflow and group these lists per category, we can count the number of occurrences of each dataset title per category.

Looking at the dataset titles per category sorted by ascending number of occurrences, we observed that the correct matches for the datasets' references using a specific characteristic feature were always among the top 6 items.

7. Conclusion and future work

We have presented an approach for identifying references to datasets in social sciences papers. It works in real time and does not require any training dataset. There are just some manual tasks in the approach such as initially cleaning the dictionary of abbreviations, or making final decisions among multiple candidates suggested for the datasets cited by the given paper. We have achieved an F-measure of 0.84 for the detection task and an F-measure of 0.83 for finding correct matches for each reference in the gold standard. Although the da|ra registry is large and it is growing fast, there are still many datasets that have not yet been registered there. This circumstance will adversely affect the task of detecting references to datasets in papers and matching them to items in da|ra. After the evaluation, our observations reveal that da|ra could cover only 64 percent of datasets in our test corpus.

Future work will focus on improving the accuracy of detecting references to the datasets supported so far, and on extending the coverage to all datasets. Accuracy can be improved by better similarity metrics, e.g., taking into account synonyms and further metadata of datasets in addition to the title. Other algorithms such as identifying the central dataset(s) on which a paper is based can improve the ranked list generated by similarity metrics. The identification of central dataset(s) is possible after pairing a share of references of datasets in a given paper with titles in da|ra, and then this identification affects the ranking of rest of the references.

Coverage can be improved by taking into account further datasets, which are not registered in da|ra. One promising further source of datasets is OpenAIRE, the Open Access Infrastructure for Research in Europe, which so far covers more than 16,000 datasets from all domains inluding social science but is rapidly growing thanks to the increasing attention paid to open access publishing in the EU. The OpenAIRE metadata can be consumed via OAI-PMH, or, in an even more straightforward way, as linked data (cf. our previous work, Vahdati et al. [27]). For each dataset reference in the paper, we will model the precise position of that reference, and the algorithm's confidence in each possible matching dataset. In a mid-term perspective, solutions for identifying dataset references in papers that have been published already could be made redundant by a wider adoption of standards for properly citing datasets while authoring papers, and corresponding tool support for authors.

Acknowledgements This work has been partly funded by the DFG project "Opening Scholarly Communication in Social Sciences" (grant agreements SU 647/19-1 and AU 340/9-1), and by the European Commission under grant agreement 643410. We thank Katarina Boland from the InFoLiS II project (MA 5334/1-2) for helpful discussions and for generating the gold standard for our evaluation. This paper is an extension of our previous work [28]. The implementation code is available in the OSCOSS repository (link: `github.com/OSCOSS/Dataset_Detcter/tree/master/src`).

References

[1] Brigitte Mathiak and Katarina Boland. Challenges in matching dataset citation strings to datasets in social science. *D-Lib Magazine*, 21(1/2), 2015. doi: 10.1045/january2015-mathiak.

[2] Sunita Sarawagi. Information extraction. *Foundations and Trends in Information Retrieval in Databases*, 1(3):261–377, 2007. doi: 10.1561/1900000003.

[3] Meiyu Lu, Srinivas Bangalore, Graham Cormode, Marios Hadjieleftheriou, and Divesh Srivastava. A dataset search engine for the research document corpus. In *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, pages 1237–1240. IEEE, 2012. doi: 10.1109/ICDE.2012.80.

[4] Micah Altman and Gary King. A proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, 13(3/4), 2007. doi: 10.1045/march2007-altman.

[5] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007. doi: 10.1075/li.30.1.03nad.

[6] SJ Pepler and K O'Neil. Preservation intent and collection identifiers: Claddier project report ii, 2008. URL `http://purl.org/net/epubs/work/43640`.

[7] Allen H Renear, Simone Sacchi, and Karen M Wickett. Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.

[8] Gerard Salton and Christopher Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[9] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

[10] Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In *ICML 97 Proceedings of the Fourteenth International Conference on Machine Learning*, pages 143–151. Morgan Kaufmann, 1997.

[11] David M W Powers. *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation*, volume 2. Journal of Machine Learning Technologies, 2011.

[12] Sungjick Lee and Han-joon Kim. News keyword extraction for topic tracking. In *Networked Computing and Advanced Information Management, 2008 (NCM 08)*, volume 2, pages 554–559. IEEE, 2008.

[13] Peter D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL 01, pages 491–502. Springer-Verlag, 2001. URL `http://dl.acm.org/citation.cfm?id=645328.650004`.

[14] Ayush Singhal and Jaideep Srivastava. Data extract: Mining context from the web for dataset extraction. *International Journal of Machine Learning and Computing*, 3(2), 2013. doi: 10.7763/IJMLC.2013.V3.306.

[15] Christoph Schaefer, Daniel Hienert, and Thomas Gottron. Normalized relevance distance -– a stable metric for computing semantic relatedness over reference corpora. In *European Conference on Artificial Intelligence (ECAI)*, volume 263, pages 789–794, 2014. doi: 10.3233/978-1-61499-419-0-789.

[16] Mehran Sahami and Timothy D. Heilman. *A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets*. WWW 06. ACM, 2006. doi: 10.1145/1135777.1135834.

[17] Muhammad Tanvir Afzal, Hermann Maurer, Wolf-Tilo Balke, and Narayanan Kulathuramaiyer. *Rule Based Autonomous Citation Mining with Tierl*. Journal of Digital Information Management 8(3) 196-204, 2010.

[18] Kuo Zhang, Hui Xu, Jie Tang, and Juanzi Li. Keyword extraction using support vector machine. In *Proceedings of the 7th International Conference on Advances in Web-Age Information Management*, WAIM 06, pages 85–96. Springer-Verlag, 2006. doi: 10.1007/11775300_8.

[19] Hui Han, C. L. Giles, E. Manavoglu, Hongyuan Zha, Zhenyue Zhang, and E. A. Fox. *Automatic Document Metadata Extraction Using Support Vector Machines*. ACM/IEEE 2003 Joint Conference on Digital Libraries, 2003. doi: 10.1109/JCDL.2003.1204842.

[20] Jasmeen Kaur and Vishal Gupta. Effective approaches for extraction of keywords. *IJCSI International Journal of Computer Science*, 7(6):144–148, 2010.

[21] Chengzhi Zhang, Huilin Wang, Yao Liu, Dan Wu, Yi Liao, and Bo Wang. Automatic keyword extraction from documents using conditional random fields. *Computational and Information Systems*, 4(3), 2008.

[22] Bin-Ge Cui and Xin Chen. An improved hidden markov model for literature metadata extraction. In *6th International Conference on Intelligent Computing, ICIC 2010, Changsha, China*, 2010. doi: 10.1007/978-3-642-14922-1_26.

[23] Francis Kubala, Richard Schwartz, Rebecca Stone, and Ralph Weischedel. Named entity extraction from speech. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 287–292, 1998.

[24] Simone Marinai. Metadata extraction from PDF papers for digital library ingest. In *Document Analysis and Recognition, 2009. ICDAR 09. 10th International Conference on*, pages 251–255. IEEE, 2009. doi: 10.1109/ICDAR.2009.232.

[25] R. Kern, K. Jack, M. Hristakeva, and M. Granitzer. *Teambeam-Meta-Data Extraction from Scientific Literature*. D-Lib Magazine 18(7/8), 1, 2012. doi: 10.1045/july2012-kern.

[26] Katarina Boland, Dominique Ritze, Kai Eckert, and Brigitte Mathiak. Identifying references to datasets in publications. In *Proceedings of the Second International Conference on Theory and Practice of Digital Libraries (TDPL 2012)*, pages 150–161. Springer, 2012.

[27] Sahar Vahdati, Farah Karim, Jyun-Yao Huang, and Christoph Lange. Mapping large scale research metadata to linked data: A performance comparison of HBase, CSV and XML. In *Metadata and Semantics Research*, Communications in Computer and Information Science. Springer, 2015. doi: 10.1007/978-3-319-24129-6_23.

[28] Behnam Ghavimi, Philipp Mayr, Sahar Vahdati, and Christoph Lange. Identifying and improving dataset references in social sciences full-–texts. *International Conference on Electronic Publishing (ELPUB)*, Göttingen, Germany, 7-–9 June 2016. URL `http://arxiv.org/abs/1603.01774`.