Scientometrics
https://doi.org/10.1007/s11192-020-03382-z

unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata

Tarek Saier[1] · **Michael Färber**[1]

Received: 30 September 2019

Abstract
In recent years, scholarly data sets have been used for various purposes, such as paper recommendation, citation recommendation, citation context analysis, and citation context-based document summarization. The evaluation of approaches to such tasks and their applicability in real-world scenarios heavily depend on the used data set. However, existing scholarly data sets are limited in several regards. In this paper, we propose a new data set based on all publications from all scientific disciplines available on arXiv.org. Apart from providing the papers' plain text, in-text citations were annotated via global identifiers. Furthermore, citing and cited publications were linked to the Microsoft Academic Graph, providing access to rich metadata. Our data set consists of over one million documents and 29.2 million citation contexts. The data set, which is made freely available for research purposes, not only can enhance the future evaluation of research paper-based and citation context-based approaches, but also serve as a basis for new ways to analyze in-text citations, as we show prototypically in this article.

Keywords Scholarly data · Citations · arXiv.org · Digital libraries · Data set

Introduction

A variety of tasks use scientific paper collections to help researchers in their work. For instance, research paper recommender systems have been developed (Beel et al. 2016). Related are systems that operate on a more fine-grained level within the full text, such as the textual contexts in which citations appear (i.e., citation contexts). Based on citation contexts, things like the citation function (Teufel et al. 2006a, b; Moravcsik and Murugesan 1975), the citation polarity (Ghosh et al. 2016; Abu-Jbara et al. 2013), and the citation importance (Valenzuela et al. 2015; Chakraborty and Narayanam 2016) can be determined. Furthermore, citation contexts are necessary for context-aware citation recommendation (He et al. 2010; Ebesu and Fang 2017), as well as for citation-based document

✉ Tarek Saier
tarek.saier@kit.edu

Michael Färber
michael.faerber@kit.edu

[1] Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Published online: 02 March 2020

summarization tasks (Chandrasekaran et al. 2019), such as citation-based automated survey generation (Mohammad et al. 2009) and automated related work section generation (Chen and Zhuge 2019).

The evaluation of approaches developed for all these tasks as well as the actual applicability and usefulness of developed systems in real-world scenarios heavily depend on the used data set. Such a data set is typically a collection of papers provided in full text, or a set of already extracted citation contexts, consisting of, for instance, 1–3 sentences each. Existing data sets, however, do not fulfill all of the following criteria (see section "Existing data sets" for more details):

1. *Size.* The data set can be comparatively small (below 100,000 documents) which makes it difficult to use it for training and testing machine learning approaches;
2. *Cleanliness.* The papers' full texts or citation contexts are often very noisy due to the conversion from PDF to plain text and due to encoding issues;
3. *Global citation annotations.* No links from the citations in the text to the structured representations of the cited publications across documents are provided;
4. *Data set interlinkage.* Data sets often do not provide identifiers of the citing and cited documents from widely used bibliographic databases, such as DBLP[1] or the Microsoft Academic Graph[2] (MAG);
5. *Cross-domain coverage.* Often, only a single scientific discipline is available for evaluating or applying an approach to a paper or citation-based task.

In this paper we propose a new scholarly data set, which we call *unarXive*.[3] The data set is built for tasks based on papers' full texts, in-text citations, and metadata. It is freely available at http://doi.org/10.5281/zenodo.3385851 and the implementation for creating it at https://github.com/IllDepence/unarXive.

Table 1 gives an overview of the proposed data set. Note that throughout this article, we refer to links between publications on the *document level* as "references" (corresponding to entries in a section "bibliography" or "references" near the end of a document), whereas on the *text level* we speak of "citations" (indicated by markers within the text associated with a reference). The proposed data set consists of over one million full text documents (about 269 million sentences) and links to 2.7 million unique publications via 15.9 million unique references and 29.2 million citations. Thus, we argue that it is considerably large, fulfilling criterion (1). By using publications' LaTeX source files and developing a highly accurate transformation method that converts LaTeX to plain text, we can resolve issue (2). Besides the pure papers' content, in-text citations are annotated directly in the text via global identifiers, thereby covering aspect (3). As far as possible, (citing and cited) documents are linked to the Microsoft Academic Graph (Sinha et al. 2015) (cf. item (4)). This enables us to use the arXiv paper content in combination with the metadata in the MAG, which, as of February 2019, contains data on 213 million publications along with metadata about researchers, venues, and fields of study. Our data set also fulfills constraint (5) as all

[1] See https://dblp.uni-trier.de/.

[2] See https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/ and http://ma-graph.org.

[3] The name is derived from the source name *arXiv* and the verb *to unarchive*, indicating the extraction of files from an archive.

disciplines covered in arXiv are included. This enables researchers to analyze papers from several disciplines and to compare approaches using scholarly data across disciplines.

Considering the application of our data set, we argue, that it not only can be used as a new large data set for evaluating paper-based and citation-based approaches with unlimited citation context lengths (since the publications' full texts are available), but also be a basis for novel ways of paper analytics within bibliometrics and scientometrics. For instance, based on the citation contexts and the citing and cited papers' metadata in the MAG, analyses on biases in the writing and citing behavior of researchers—e.g. related to authors' affiliation (Reingewertz and Lutmar 2018) or documents' language (Liang et al. 2013; Liu et al. 2018)—can be performed. Furthermore, (sophisticated) deep learning approaches, as they are also widely used in the digital library domain recently (Ebesu and Fang 2017), require huge amounts of training data. Our data set allows to overcome this hurdle and investigate how far deep learning approaches can lead us. Overall, we argue that with our data set we can significantly bring the state of the art of big scholarly data one step forward.

We make the following contributions in this paper:

1. We propose a large, interlinked scholarly data set with papers' full texts, annotated in-text citations, and links to rich metadata. We describe its creation process in detail and provide both the data as well as the creation process implementation to the public.
2. We manually evaluate the validity of our reference links on a sample of 300 references, thereby providing insight into our citation network's quality.
3. We calculate statistical key figures and analyze the data set with respect to its contained references and citations.
4. We compare our reference links to those in the MAG, and manually evaluate the validity of links only appearing in either of the data sets. In doing so, we identify a large number of documents where the MAG lacks coverage.
5. We analyze the likelihood with which in-text citations in our data set refer to specific parts of a cited document depending on the discipline of the citing *and* cited document. Such an analysis is only possible with word level precision citation marker positions annotated in full text *and* metadata on citing as well as cited documents. The analysis therefore can showcase the practicability of our data set.

The paper is structured as follows: After outlining related data sets in section "Existing data sets", we describe in section "Data set creation" how we created our data set. This is followed by statistics and key figures in section "Statistics and key figures". In section "Evaluation of citation data validity and coverage", we evaluate the validity and coverage of our reference links. Section "Analysis of citation flow and citation contexts" is dedicated to the analysis of the citation flow and the contexts within our data set. We conclude in section "Conclusion" with a summary and an outlook.

Existing data sets

Table 2 gives an overview of related data sets. CiteSeerX can be regarded as the most frequently used evaluation data set for citation-based tasks. For our investigation, we use the snapshot of the entire CiteSeerX data set as of October 2013, published by Huang et al. (2015). This data set consists of 1,017,457 papers, together with 10,760,318 automatically extracted citation contexts. This data set has the following drawbacks (Roy et al. 2016;

Table 1 Overview of the proposed data set

	Citing documents	References		Cited documents
		Outgoing	Incoming	
Full data set	**1,043,126**	15,954,664	15,954,664	**2,746,288**
Full text	1,043,126	15,954,664	7,181,576	736,597
Linked to MAG	994,351	15,846,351	15,954,664	2,746,288
By discipline				
Physics	662,894	9,300,576	7,827,072	921,852
Mathematics	237,422	3,426,117	5,062,033	906,301
Computer science	111,694	2,526,656	1,876,401	425,860
Other	31,116	701,315	1,189,158	492,275

Data: http://doi.org/10.5281/zenodo.3385851

Code: https://github.com/IllDepence/unarXive

Färber et al. 2018): The provided meta-information about cited publications is often not accurate. Citing and cited documents are not interlinked to other data sets. Moreover, the citation contexts can contain noise from non-ASCII characters, formulas, section titles, missed references and/or other "unrelated" references, and do not begin with a complete word.

The PubMed Central Open Access Subset is another large data set that has been used for citation-based tasks (Gipp et al. 2015; Duma et al. 2016; Galke et al. 2018). Contained publications are already processed and available in the JATS (Huh 2014) XML format. While the data set overall is comparatively clean, heterogeneous annotation of citations within the text and mixed usage of identifiers of cited documents (PubMed, MEDLINE, DOI, etc.) make it difficult to retrieve high quality citation interlinkings of documents from the data set[4] (Gipp et al. 2015).

Beside the abovementioned, there are other collections of scientific publications. Among them are the ACL Anthology corpus (Bird et al. 2008) and Scholarly Dataset 2 (Sugiyama and Kan 2015). Note that these data sets only contain the publications themselves, typically in PDF format. Therefore, using such data sets for paper-based or citation-based approaches is troublesome, since one must preprocess the data (i.e., (1) extract the content without introducing too much noise, (2) specify global identifiers for cited papers, and (3) annotate citations with those identifiers). Furthermore, there are data sets for evaluating paper recommendation tasks, such as CiteULike[5] or Mendeley.[6] These, however, only provide metadata about publications or are not freely available for research purposes.

Prior to publishing the data set described in this paper, we already published a data set with annotated arXiv papers' content in the past (Färber et al. 2018). In comparison, our new data set is superior to this initial version in the following regards:

[4] To be more precise, the heterogeneity makes the usage of the data set *as is* unfeasible. Resolving references to a single consistent set of identifiers retrospectively would be an option, but comparatively challenging in the case of PubMed, because of the frequent usage of special notation in publication titles; see also: http://www.sciplore.org/files/citrec/CITREC_Parser_Documentation.pdf.

[5] Hosted at http://citeulike.org/ until March 2019.

[6] See https://data.mendeley.com/.

Table 2 Overview of existing data sets

Data set	#P.	Cit. cont.	Scope	Full text	Ref. IDs
CiteSeerX (Caragea et al. 2014) / RefSeer (Huang et al. 2015)	1 M	400 chars	(all)	No	No
PubMed Central OAS[a]	2.3 M	extractable	BM/LS	Yes	Mixed
Scholarly Dataset 2 (Sugiyama and Kan 2015)	100 k	extractable*	CS	Yes	No
arXiv CS (Färber et al. 2018)	90 k	1 entence	CS	Yes	DBLP
ACL-ARC (Bird et al. 2008)	11 k	extractable*	CS/CL	Yes	No
ACL-AAN (Radev et al. 2013)	18 k	extractable*	CS/CL	Yes	No

[a] See https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

P.= Number of papers; Cit. cont. = Citation contexts; Ref. IDs = Reference IDs; CS = Computer Science, BM = Biomedicine; LS = Life Sciences; CL = Computer Linguistics; *extractable** indicates that extraction might be error-prone due to papers only being available in PDF format

1. The new data set is considerably larger (1 M instead of 90 k documents).
2. The new data set provides a similar level of cleanliness to the old data set regarding the papers' full texts and citation contexts.
3. A new method for resolving references to consistent global identifiers has been developed. Contrary to the old method, the new method has been evaluated and performs very well (see section "Citation data validity").
4. While the old data set links documents solely to DBLP, which covers computer science papers, the new data set links documents to the Microsoft Academic Graph, which covers all scientific disciplines and which has been used frequently in the digital library domain in recent years (Mohapatra et al. 2019).
5. While the old data set is restricted to computer science, the new data set covers all domains of arXiv (see section "Statistics and key figures" and Fig. 7).

Lastly, compared to the initial publication of our new data set (Saier and Färber 2019), this journal article provides significantly more details and insights into the data set's creation process (see section "Data set creation") and its resulting characteristics (see sections "Evaluation of citation data validity and coverage" and "Analysis of citation flow and citation contexts"). Moreover, the data set has been further improved. Most notably, while in the initial version, only citing papers were associated with arXiv identifiers and only cited papers had been linked to the MAG, we now provide both types of IDs for both sides. This means, that for nearly all documents, MAG metadata is easily accessible, and full text is not only available for all citing papers but now also for over a quarter of the cited papers.

Data set creation

Scientific publications are usually distributed in formats targeted at *human consumption* (e.g., PDF) or, in cases like arXiv, also as source files the aforementioned (e.g., LaTeX sources for generating PDFs). Citation-based tasks, such as context-aware citation recommendation, in contrast, require *automated processing* of the publications' textual contents as well as the documents' interlinking through in-text citations. The creation of a data set for such tasks therefore encompasses two main steps: extraction of plain text and resolution

of references. In the following, we will describe how we approached these two steps using arXiv publications' LaTeX sources and the Microsoft Academic Graph.

Used data sources

The following two resources are the basis of the data set creation process.

arXiv hosts over 1.5 million documents from August 1991 onward.[7] They are available not only as PDF, but (in most cases) also as LaTeX source files. The discipline most prominently represented is physics, followed by mathematics, with computer science seeing a continued increase in percentage of submissions ranking third (see Fig. 7). The availability of LaTeX sources makes arXiv documents particularly well suited for extracting high quality plain text and accurate citation information. So much so, that it has been used to generate ground truths for the evaluation of PDF-to-text conversion tools (Bast and Korzen 2017).

Microsoft Academic Graph is a very large, automatically generated data set on 213 million publications, related entities (authors, venues, etc.), and their interconnections through 1.4 billion references.[8] It has been widely used as a repository of all publications in academia in the fields of bibliometrics and scientometrics (Mohapatra et al. 2019). While pre-extracted citing sentences are available, these do not contain annotated citation marker positions. Full text documents are also not available. The size of the MAG makes it a good target for matching reference strings[9] against it, especially given that arXiv spans several disciplines.

Pipeline overview

To create the data set, we start out with arXiv sources (see Fig. 1). From these we generate, per publication, a plain text file with the document's textual contents and a set of database entries reflecting the document's reference section. Association between reference strings and in-text citation locations are preserved by placing citation markers in the text. In a second step, we then iterate through all reference strings in the database and match them against paper metadata records in the MAG. This gives us full text arXiv papers with (word level precision) citation links to MAG paper IDs. As a final step, we enrich the data with MAG IDs on the citing paper side (in addition to the already present arXiv IDs) and arXiv IDs on the cited paper side (in addition to the already present MAG IDs)—this is a straightforward process, because the paper metadata in the MAG includes source URLs, meaning papers found on arXiv have an arXiv.org source URL associated with them, such that a mapping from arXiv IDs to MAG IDs can be created.

Listing 2 shows how our data set looks like. In the following, we describe the main steps of the data set creation process in more detail.

[7] See https://arxiv.org/stats/monthly_submissions.

[8] Numbers as of February 2019.

[9] I.e., the entries in the reference section of a publication. See Lst. 1 for examples.

LaTeX parsing

In the following, we will describe the tools considered for parsing LaTeX, the challenges we faced in general and with regard to arXiv sources in particular, and our resulting approach.

Tools

We took several tools for a direct conversion from LaTeX to plain text or to intermediate formats into consideration and evaluated them. Table 3 gives an overview of our results. Half of the tools failed to produce any output for a large amount of arXiv documents we used as test input and were therefore deemed not robust enough. *GrabCite* (Färber et al. 2018) is able to parse 78.5% of arXiv CS documents but integrates resolving references (see section "Resulting approach") against DBLP into the parsing process and therefore would require significant modification to fit our new system architecture. *LaTeXML* and *Tralics* are both robust and can be used as LaTeX conversion tools as is. Based on subsequent tests, we observed that *LaTeXML* needs on average 7.7 s (3.3 if formula environments are heuristically removed beforehand) to parse an arXiv paper, while *Tralics* needs 0.09. Because the quality of their output seemed comparable, we chose to use *Tralics*.

Challenges

Apart from the general difficulty of parsing LaTeX due to its feature richness and people's free-spirited use of it, we especially note difficulty in dealing with extra packages not included in documents' sources.[10] While *Tralics*, for example, is supposed to deal with *natbib* citations,[11] normalization of such citations leads to a decrease of citation markers not being able to be matched to an entry in the document's reference section from 30 to 5% in a sample of 565,613 citations we tested.

Resulting approach

Our LaTeX parsing solution consists of three steps: flattening, parsing, and output generation. First, we flatten each arXiv document's sources to a single LaTeX file using *latexpand*[12,13] and normalize citation commands (e.g. `\citep *`, `\citet[see]`, `\citealt`, etc. to `\cite`) to prevent parsing problems later on. In the second step, we then generate an XML representation of the LaTeX document using *Tralics*. Lastly, we go through the generated XML structure and produce two types of output—(i) an annotated plain text file with the document's textual contents and (ii) database entries reflecting the document's reference section. For (i) we replace XML nodes that represent formulas, figures, tables, as well as intra-document references with replacement tokens and turn XML nodes

[10] The arXiv guidelines specifically suggest the omission of such (see https://arxiv.org/help/submit_tex#wegotem).

[11] See https://www-sop.inria.fr/marelle/tralics/packages.html#natbib.

[12] See https://ctan.org/pkg/latexpand.

[13] We also tested flatex (https://ctan.org/pkg/flatex) and flap (https://github.com/fchauvel/flap) but got the best results with latexpand.

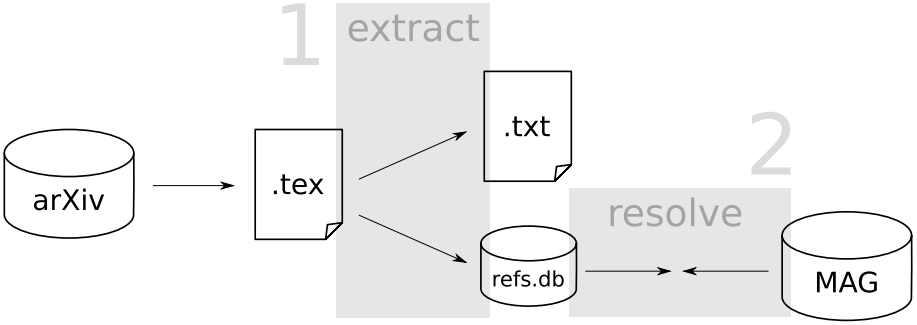

Fig. 1 Schematic representation of the data set generation process

Table 3 Comparison of tools for parsing LaTeX

Tool	Output	Robust	Usable as is
plastex[a]	DOM	No	Yes
TexSoup[b]	Document tree	No	Yes
opendetex[c] /detex[d]	Plain Text	No	Yes
GrabCite (Färber et al. 2018)	Plain text + resolved ref.	Yes	No
LaTeXML[e]	XML	Yes	Yes
Tralics[f]	XML	Yes	Yes

[a]See https://github.com/tiarno/plastex.

[b]See https://github.com/alvinwan/texsoup.

[c]See https://github.com/pkubowicz/opendetex.

[d]See https://www.freebsd.org/cgi/man.cgi?query=detex.

[e]See https://github.com/brucemiller/LaTeXML.

[f]See https://www-sop.inria.fr/marelle/tralics/.

originating from citation markers in the LaTeX source (i.e., `\cite`) into plain text citation annotation markers. For (ii), each entry in the document's reference section is assigned a unique identifier, its text is stored in a database, and the identifier put into the corresponding annotation in the plain text (cf. Listing 2).

Reference resolution

Resolving references to globally consistent identifiers (e.g. detecting that the reference strings (1), (2), and (3) in Listing 1 all reference the same document) is a challenging and still unsolved task (Nasar et al. 2018). Given it is the most distinctive singular part of a publication, we base our reference resolution on the title of the cited work and use other pieces of information (e.g., the authors' names) only in secondary steps. In the following, we will describe the challenges we faced, matching arXiv documents' reference strings against MAG paper records, and how we approached the task.

```
(1) V. N. Senoguz and Q. Shafi, arXiv:hep-ph/0412102
(2) V.N. Senoguz and Q. Shafi, Phys. Rev. D 71 (2005) 043514.
(3) V. N. Senoguz and Q. Shafi, ''Reheat temperature in supersymme
    tric hybrid inflation models,'' Phys. Rev. D 71, 043514 (2005)
     [hep-ph/0412102].
(4) V.Sauli, JHEP 02, 001 (2003).
(5) Aaij, Roel, et al. "Search for the $B^{0}_{s} \to \eta^{\prime
    }\phi$ decay" Journal of High Energy Physics 2017.5 (2017): 15
    8.
(6) According to the numerous discussions with my colleagues <remo
    ved> and <removed> an experimental verification of our theoret
    ical predictions is feasible.
```

Listing 1 Examples of reference strings

Challenges

Reference resolution can be challenging when reference strings contain only minimal amounts of information, when formulas or other special notation is used in titles, or when they refer to non publications (e.g., Listing 1, (4)–(6)). Another problem we encountered was noise in the MAG. One such case are the MAG papers with IDs `2167727518` and `2763160969`. Both are identically titled *"Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC"* and dated to the year 2012. But while the former is cited 17k times and cites 112 papers within the MAG, the latter is a neither cited nor cites any other papers.[14] Taking the number of citations into account when matching references, reduced the number of mismatches in this particular case from 2,918 to 0 and improved the overall quality of matches in general.

Resulting approach

Our reference resolution procedure can be broken down in two steps: title identification and matching. If contained in the reference string, title identification is performed based on an arXiv ID or DOI (where we retrieve the title from an arXiv metadata dump or via crossref.org[15]); otherwise we use Neural ParsCit (Prasad et al. 2018).[16] The identified title is then matched against the normalized titles of all publications in the MAG. Resulting candidates are considered, if at least one of the author's names (as given in the MAG) is present in the reference string. If multiple candidates remain, we judge by the citation count given in the MAG—this particularly helps mitigate matches to rouge almost-duplicate entries in the MAG, which often have few to no citations, like paper `2763160969` mentioned in the previous section.

[14] The MAG record with ID `2763160969` appears to be a noisy duplicate caused by a web source with easily misinterpretable author information (only a partial list is displayed).

[15] See https://www.crossref.org/.

[16] For title identification we also considered two other state of the art (Tkaczyk et al. 2018) tools, namely CERMINE (Tkaczyk et al. 2015) and GROBID (Lopez 2009). However, we found CERMINE to be considerably slower than the other tools. And while GROBID showed comparable speed and output quality in preliminary tests, Neural ParsCit's tag based output format was more straightforward to integrate than the faceted TEI format structures that GROBID's reference parser module returns.

```
It has over 79 million images stored at the resolution of FORMULA
. Each image is labeled with one of the 75,062 non-abstract nouns
in English, as listed in the Wordnet{{cite:9ad20b7d-87d1-47f5-aeed
-10a1cf89a2e2}}{{cite: 298db7f5-9ebb-4e98-9ecf-0bdda28a42cb}} lexi
cal database.
-----------------------------------------------------------------
[uuid]         [citing..] [cited..]   ...  [reference_string]
9ad20b7d-87d1  1412.3684  2081580037  ...  George A. Miller (1995)
-47f5-aeed-..                              . WordNet: A Lexical ..
298db7f5-9ebb  1412.3684  2038721957  ...  Christiane Fellbaum (19
-4e98-9ecf-..                              98), ""WordNet: An El..
-----------------------------------------------------------------
[paperid]   [originaltitle]                              [publ..]  ..
2038721957  WordNet : an electronic lexical database      MIT Press ..
2081580037  WordNet: a lexical database for English       ACM       ..
-----------------------------------------------------------------
2131463865|2038721957|2081580037|1412.3684|||It has over 79 millio
n images stored at the resolution of FORMULA . Each image is label
ed with one of the 75,062 non-abstract nouns in English, as listed
 in the Wordnet CIT MAINCIT lexical database. It has been noted th
at many of the labels are not reliable CIT .
```

Listing 2 Excerpts from (top to bottom) a paper's plain text, corresponding entries in the references database, entries in the MAG, and extracted citation context CSV

Result format

Listing 2 shows some example content from the data set. In addition to the paper plain text files and the references database, we also provide the citation contexts of all successfully resolved references extracted to a CSV file as well as a script to create custom exports.[17] For the provided CSV export, we set the citation context length to 3 sentences—the sentence containing the citation as well as the one before and after—as used by Tang et al. (2014) and Huang et al. (2015). Each line in an export CSV has the following columns: cited MAG ID, adjacent cited MAG IDs, citing MAG ID, cited arXiv ID, adjacent cited arXiv IDs, citing arXiv ID, text (see bottom of Listing 2). Citations are deemed adjacent, if they are part of a citation group or are at most 5 characters apart (e.g. "*[27,42]*", *"[27], [42]"* or *"[27] and [42]"*). The IDs of adjacent cited documents are added, because those documents are cited in an almost identical context (i.e. only a few characters to the left or right).

Statistics and key figures

In this section we present the data set and its creation process in terms of numbers. Furthermore, insight into the distribution of references and citation contexts is given.

[17] See Python script `extract_contexts.py` bundled with the data set for details.

Creation process

We used an arXiv source dump containing all documents up until the end of 2018 (1,492,923 documents). 114,827 of these were only available in PDF format, leaving 1,378,096 sources. Our pipeline output 1,283,584 (93.1%) plain text files, 1,139,790 (82.7%) of which contained citation markers. The number of reference strings identified is 39,694,083, for which 63,633,427 citation markers were placed within the plain text files. This first part of the process took 67 h to run, unparallelized on a 8 core Intel Core i7-7700 3.60 GHz machine with 64 GB of memory.

Of the 39,694,083 reference strings, we were able to match 16,926,159 (42.64%) to MAG paper records. For 31.32% of the reference strings we could neither find an arXiv ID or DOI, nor was Neural ParsCit able to identify a title.[18] For the remaining 26.04% a title was identified, but could not be matched to the MAG. Of the matched 16.9 million items' titles, 52.60% were identified via Neural ParsCit, 28.31% by DOI and 19.09% by arXiv ID. Of the identified DOIs, 32.9% were found as is, while 67.1% were heuristically determined. This was possible because the DOIs of articles in journals of the American Physical Society follow predictable patterns. The matching process took 119 h, run in 10 parallel processes on a 64 core Intel Xeon Gold 6130 2.10 GHz machine with 500 GB of memory.

Comparing the performance of our approach using all papers (1991–2018) to using only the papers from 2018 (i.e. recent content), we note that the percentage of successfully extracted plain texts goes up from 93.1 to 95.9% (82.7 to 87.8% only counting plain text files containing citation markers) and the percentage of successfully resolved references increases from 42.64 to 59.39%. A possible explanation for the latter would be, that there is more and higher quality metadata coverage (MAG, crossref.org, etc.) of more recent publications.

Resulting data set

Our data set consists of *2,746,288 cited papers, 1,043,126 citing papers, 15,954,664 references and 29,203,190 citation contexts*.[19]

Figure 2 shows the number of citing documents for all cited documents. There is one cited document with over 10,000 citing documents, another 8 with more than 5,000 and another 14 with more than 3,000. 1,485,074 (54.07%) of the cited documents are cited at least two times, 646,509 (23.54%) at least five times. The mean number of citing documents per cited document is 5.81 (SD 28.51). Figure 3 shows the number of citation contexts per entry in a document's reference section. 10,537,235 (66.04%) entries have only one citation context, the maximum is 278, the mean 1.83 (SD 2.00).

[18] To assess whether or not the large percentage of reference strings without identified title is due to Neural ParsCit missing a lot of them, we manually check its output for a random sample of 100 papers (4027 reference strings). We find that 99% of cases with no title identified actually do not contain a title—like for example items (1), (2) and (4) in Lst. 1. These kind of references seem to be most common in physics papers. The 1% where a title was missed were largely references to non-English titles and books. We therefore conclude that the observed numbers largely reflect the actual state of reference strings rather than problems with the approach taken.

[19] References that were successfully matched to a MAG record but have no associated citation markers (due to parsing errors; cf. section "Challenges") are not counted here.

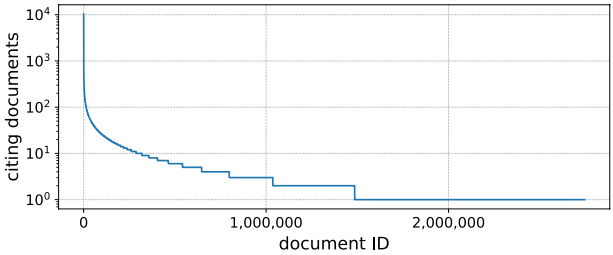

Fig. 2 Number of citing documents per cited document

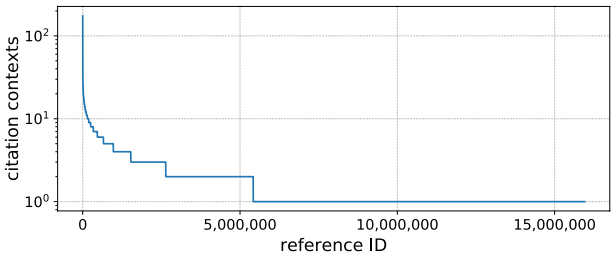

Fig. 3 Number of citation contexts per reference

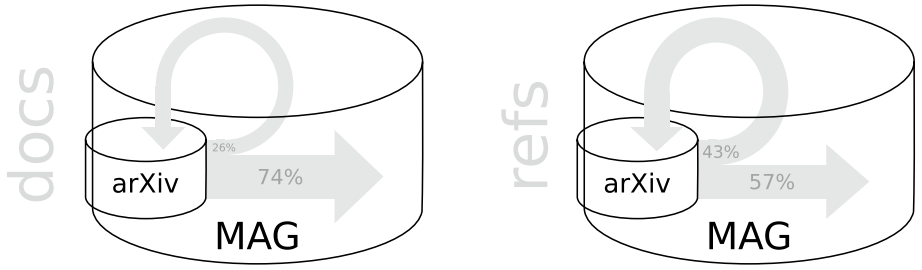

Fig. 4 Visualization of the citation flow in terms of documents and references from arXiv to the MAG

Because not all documents referenced by arXiv papers are hosted on arXiv itself, we additionally visualize the citation flow with respect to the MAG in Fig. 4. 95% of our citing documents are contained in the MAG. Of the cited documents, 26% are contained in arXiv and therefore included as full text, while 74% are only included as MAG IDs. On the level of references, this distribution shifts to 43/57. The high percentages of citation links contained within the data set can be explained due to the fact, that in physics and mathematics—which make up a large part of the data set—it is common to self-archive papers on arXiv.

Evaluation of citation data validity and coverage

Citation data validity

To evaluate the validity of our reference resolution results, we take a random sample of 300 matched reference strings and manually check for each of them, if the correct record in the MAG was identified. This is done by viewing the reference string next to the matched

Table 4 Confidence intervals for a sample size of 300 with 297 positive results as given by Wilson score interval and Jeffreys interval (Brown et al. 2001)

Confidence level	Method	Lower limit	Upper limit
0.99	Wilson	0.9613	0.9975
	Jeffreys	0.9666	0.9983
0.95	Wilson	0.9710	0.9966
	Jeffreys	0.9736	0.9972

Table 5 Mismatched documents

#		Document
1	Matched	*"The Maunder Minimum"* (John A. Eddy; 1976)
	Correct	*"The Maunder Minimum: A reappraisal"* (John A. Eddy; 1983)
2	Matched	*"Support Vector Machines"* (Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani; 2013)
	Correct	*"1-norm Support Vector Machines"* (Ji Zhu, Saharon Rosset, Robert Tibshirani, Trevor J. Hastie; 2003)
3	Matched	*"The Putative Liquid-Liquid Transition is a Liquid-Solid Transition in Atomistic Models of Water"* (David Chandler, David Limmer; 2013)
	Correct	*"The putative liquid-liquid transition is a liquid-solid transition in atomistic models of water. II"* (David T. Limmer, David Chandler; 2011)

MAG record and verifying, if the former actually refers to the latter.[20] Given the 300 items, we observed 3 errors, giving us an accuracy estimate of 96% at the worst, as shown in Table 4. Table 5 shows the three incorrectly identified documents. In all three cases the misidentified document's title is contained in the correct document's title, and there is a large or complete author overlap between correct and actual match. This shows that authors sometimes title follow-up work very similarly, which leads to hard to distinguish cases.

Citation data coverage

For the 95% of our data set, where citing as well as cited document have a MAG ID, we are able to compare our citation data directly to the MAG. The composition of reference section coverage (i.e. how many of the references are reflected in each of the data sets) of all 994,351 citing documents can be seen in Fig. 5. Of the combined 26,205,834 reference links, 9,829,797 are contained in both data sets (orange), 5,918,128 are in unarXive only (blue), and 10,457,909 are in the MAG only (green). On the document level we observe, that for 401,046 documents unarXive contains more references than the MAG, and for 545,048 it is the other way around. The striking difference between reference and document level[21] suggests, that the MAG has better coverage of large reference sections. This is supported by the fact that citing papers, where the MAG contains more references, cite on average 34.28 documents, while the same average for citing papers, where unarXive contains more references, is 17.46. Investigating further, in Fig. 6 we look at the number of citing documents

[20] Further details can be found at https://github.com/IllDepence/unarXive/tree/master/doc/matching_evaluation.

[21] While the number of reference links exclusive to the MAG is about twice as high as the number of reference links exclusive to unarXive, the number of documents for which either of the data sets has better coverage is on a comparable level.

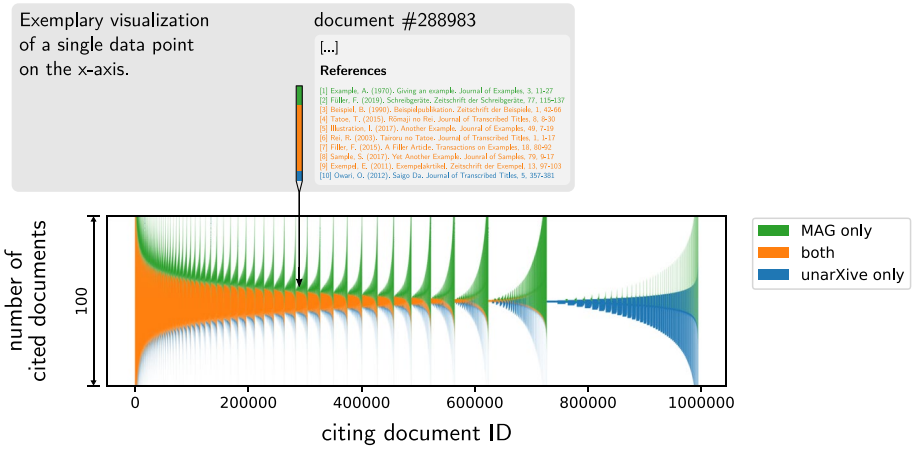

Fig. 5 Composition of reference section coverage for all citing documents (cut off at 100 cited documents)

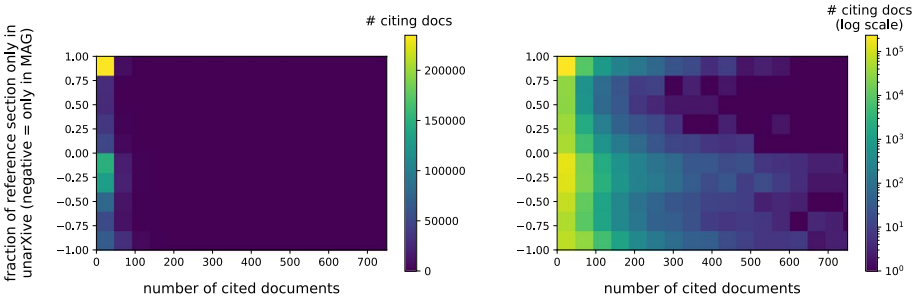

Fig. 6 Distribution of citing documents in terms of reference section size and their coverage in unarXive and MAG (cut off at 750 cited documents)

in terms of reference section *size* (x-axis) and *exclusive coverage in unarXive and MAG*[22] (y-axis). As we can see (and as the almost exclusively blue area on the right hand side of Fig. 5 suggests), there is a large number of papers, citing ≤ 50 documents, where $\geq 80\%$ of the reference section are only contained in unarXive. Put differently, there is a large portion of documents, where the reference section is covered to some degree by unarXive, but has close to no coverage in the MAG. The number of citing documents, where the MAG contains 0 references whereas unarXive has ≥ 1, is 215,291—these have an average of 15.1 references in unarXive.[23] The number of citing documents (within the 994,351 at hand), where unarXive contains 0 references whereas the MAG has ≥ 1, is 0.

[22] Calculated as $\frac{\#\text{citations only in unarXive} - \#\text{citations only in MAG}}{\#\text{citations in both} + \#\text{citations only in unarXive} + \#\text{citations only in MAG}}$.

[23] Manually looking into a sample of 100 of these documents, we find the most salient commonality to be irregularities w.r.t. to the reference section headline. 58 of the papers (55 physics, 2 quantitative biology, 1 CS) have no reference section headline, 2 have a double reference section headline and further 2 have the headline directly followed by a page break. The reason for the large number of MAG documents with no references might therefore be, that the PDF parser used can not yet deal with such cases.

Needless to say, additional references are only of value if they are valid. From both the citation links only found in unarXive, as well as those only found in the MAG, we therefore take a sample of 150 citing paper cited paper pairs and manually verify, if the former actually references the latter. This is done by inspecting the citing paper's PDF and checking the entries in the reference section against the cited paper's MAG record.[24] On the unarXive side, we observe 4 invalid links, all of which are cases similar to those showcased in Table 5. On the MAG side, we observe 8 invalid links. Some of them seem to originate from the same challenges as the ones we face, e.g. similarly titled publications by same authors, leading to misidentified *cited* papers. Other error sources are, for instance, an invalid source for a *citing* paper being used and its reference section parsed (e.g. paper ID `1504647293`, where one of the PDF sources is the third author's Ph.D. thesis instead of the described paper). Given that the citation links exclusive to unarXive appear to be half as noisy as those exclusive to the MAG, we argue that the 5,918,128 links only found in unarXive could be useful for citation and paper based tasks using MAG data. This would especially be the case for the field of physics, as it makes up a significant portion of our data set.

Analysis of citation flow and citation contexts

Because the documents in unarXive span multiple scientific disciplines, interdisciplinary analyses, such as the calculation of the flow of citations between disciplines, can be performed. Furthermore, the fact that documents are included as full text and citation markers within the text are linked to their respective cited documents, makes varied and fine grained study of citation contexts possible. To give further insight into our data set, we therefore conduct several such analyses in the following. Note that, for interdisciplinary investigations, disciplines other than physics, mathematics, and computer science are combined into *other* for space and legibility reasons, as they are only represented by a small number of publications. On the citing documents' side, these span the fields of economics, electrical engineering and systems science, quantitative biology, quantitative finance, and statistics. Combined on the cited documents' side are chemistry, biology, engineering, materials science, economics, geology, psychology, medicine, business, geography, sociology, political science, philosophy, environmental science, and art.

Citation flow

Figure 7 depicts the flow of citations by discipline for all 15.9 million matched references. As one would expect, publications in each field are cited the most from within the field itself. Notable is, that the incoming citations in mathematics are the most varied (physics and computer science combined make up 35% of the citations). As citation contexts are useful descriptive surrogates of the documents they refer to (Elkiss et al. 2008), a composition as varied as mathematics in Fig. 7 bears the question as to whether a distinction by discipline could be worth considering, when using citation contexts as descriptions of cited documents. That is, computer scientists and physicists might refer to math papers in a different way than mathematicians do. Borders between disciplines are, however, not

[24] Further details can be found at https://github.com/IllDepence/unarXive/tree/master/doc/coverage_evaluation.

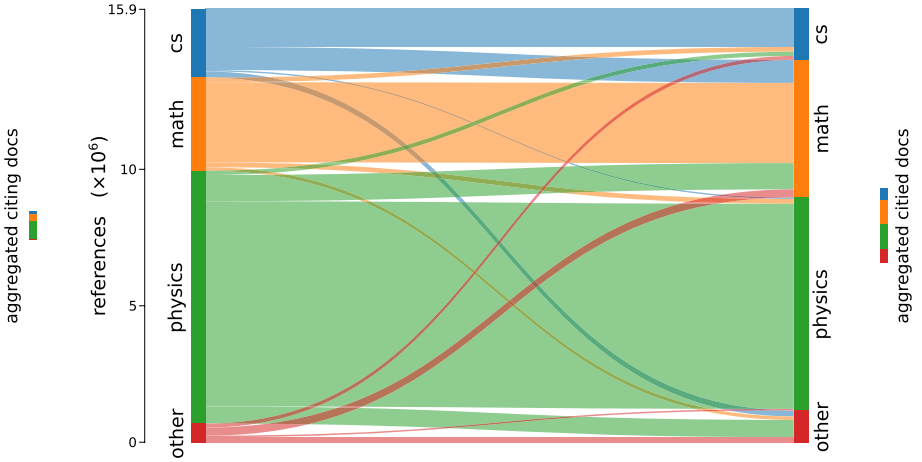

Fig. 7 Citation flow by discipline for 15.9 million references. The number of citing and cited documents per discipline are plotted on the sides

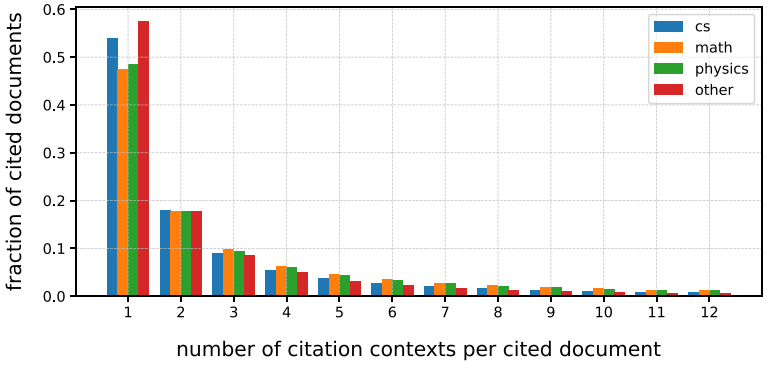

Fig. 8 Normalized distribution of the number of citation contexts per cited document

necessarily clear cut, meaning that such a distinction might not be as straight forward as the color coding in Fig. 7 suggests.

Availability of citation contexts

Another aspect that becomes relevant, when using citation contexts to describe cited documents, is the number of citation contexts available per cited publication. Figure 8 shows, that the distribution of the number of citation contexts per cited document is similar across disciplines. In each discipline, around half of the cited documents are just mentioned once across all citing documents, 17.5% exactly twice, and so on. The tail of the distribution drops a bit slower for physics and mathematics. The mean values of citation contexts per cited document are 9.5 (SD 50.3) in physics, 7.0 (SD 28.8) in mathematics, 5.1 (SD 31.1) in computer science and 3.5 (SD 11.0) for the combined other fields. This leads to two

conclusions. First, it suggests that a representation relying solely on citation contexts may only be viable for a small fraction of publications. Second, the high dispersion in the number of available citation contexts shows that means might not be very informative when it comes to citation counts aggregated over specific sets of documents.

Characteristics of citation contexts

For our analysis of the contents of citation contexts, we focus on three aspects: whether or not citations are (1) integral, (2) syntactic and (3) target section specific. These aspects were chosen, because they give particular insights into the citing behavior of researchers, as explained alongside the following definition of terms.

"Integral", "syntactic" and "target section specific" citations

We first discuss the terms *"integral"* and *"syntactic"*, which are both established in existing literature. An integral citation is one, where the name of the cited document's author appears within the citing sentence *and* has a grammatical role (Swales 1990; Hyland 1999) (e.g. "Swales [73] has argued that ..."). Similarly, a citation is syntactic, if the *citation marker* has has a grammatical role within the citing sentence (Whidby et al. 2011; Abu-Jbara and Radev 2012) (e.g. "According to [73] it is ..."). Integral citations are seen as an indication of emphasis towards the cited author (where the opposite direction would be towards the cited work) (Swales 1990; Hyland 1999). Syntactic citations are of interest, when determining how a citation relates to different parts of the citing sentence (Whidby et al. 2011; Abu-Jbara and Radev 2012). Both qualities are relevant when studying the role of citations (Färber and Sampath 2019).

Table 6 gives a more detailed account of both terms' use in literature. Note that Lamers et al. (2018) provide a classification algorithm for integral and non-integral citations that slightly differs from Swales' original definition depending on the interpretation of a citation marker's scope, but also gives a clear classification in an edge case where Swales' definition is unclear. Furthermore note, that the two ways for distinguishing syntactic and non-syntactic citations found in literature are not identical. This is in part because the method given by Abu-Jbara and Radev (2012) is kept rather simple. For the intents and purposes of our analysis we follow the definitions of Lamers et al. and Whidby et al. for *"integral"* and *"syntactic"* respectively.

As a third aspect for analysis, we define *"target section specific"* citations as those citations, where a specific section within the citation's target (i.e. the cited document) is referred to. Examples are given in Table 7. Target section specific citations are of interest for two reasons. First, in a similar fashion to integral citations, they are a particular form of citing behavior that might be used to infer characteristics of the relationship between citing author and cited document (e.g. a focus on the document rather than authors, or in depth engagement or familiarity with the cited document's contents). Second, when using citation contexts as descriptions of cited documents, such as in citation context-based document summarization, target section specific citations might benefit from special handling, as their contexts only describe a (sometimes very narrow) part of the cited document.

In the following we will analyze all three aspects (integral, syntactic, target section specific) with respect to the different scientific disciplines covered by our data set.

Table 6 Examples of citations and their categorization into integral/non-integral as well as syntactic/non-syntactic ("✓"=yes, "x"=no, "?"=unclear)

Context excerpt (*citation marker*)	Integral			Syntactic	
	Swales (1990)	Hyland (1999)	Lamers et al. (2018)	Whidby et al. (2011)	Abu-Jbara and Radev (2012)
"Swales *(1990)* has argued that ..."	✓	✓	✓	x	?
"*Swales (1990)* has argued that ..."	✓	✓	x	✓	✓
"Swales *[73]* has argued that ..."	✓	✓	✓	x	x
"Swales has argued that ...*[73]*"	✓	✓	✓	x	x
"It has been argued *(Swales, 1990)* that ..."	x	x	x	x	x
"It has been argued *[73]* that ..."	x	x	x	x	x
"According to *(Swales, 1990)* it is ..."	?	?	x	✓	✓
"According to *[73]* it is ..."	x	x	x	✓	✓
"...has been shown (see *(Swales, 1990)*)."	x	x	x	✓	x

Table 7 Examples of target section specific citations

Context excerpt (*concerns citing document* / **concerns cited document**)
"See [73], **Section 9**."
"This improves **Lemma 2** of [73], which is ..."
"Due to this, the proof is now similar to that of **Theorem 6.4** from [73]."
"The copolymer version of *Theorem 7* was derived in [73], **Theorem 3.2**."
"*Figure 1* is qualitatively similar to **Figure 3** in [73]."

Manual analysis of citation contexts

For each of the disciplines computer science, mathematics, physics, and other, we take a random sample of 300 citation contexts and manually label them with respect to being integral, syntactic, and target section specific. The result of this analysis is shown in Table 8. Each of the assigned labels is most prevalent in mathematics papers, which is furthermore true for the co-occurrence of the labels integral and syntactic. Mathematics is also the only discipline, in which citations are more likely to be syntactic than not. The difference in frequency of integral and syntactic citations might be due to variations in writing culture between the different disciplines. We think that the comparatively high frequency of target section specific citations in mathematics could be due to the fact, that in mathematics intermediate results like corollaries and lemmata are immediately reusable in related work. We further investigate target section specific citations in the following section.

Table 8 Listed per discipline is the number of citations in a sample of 300 that were labeled (1) integral, (2) syntactic, (3) simultaneously integral and syntactic, (4) target section specific

Discipline	Integral	Syntactic	Integral+syntactic	Target section specific
CS	23	88	1	5
Mathematics	48	200	13	17
Physics	12	80	2	4
Other	14	113	1	7

Automated analysis of target section specific citations

Sentences including a target section specific citation often follow distinct and predictable patterns. For example, a capitalized noun (e.g. *"Corrolary"*, *"Lemma"*, *"Theorem"*) is followed by a number and a preposition (e.g. *"in"*, *"of"*), and then followed by the citation marker (e.g. *"Corrolary 3 in [73]"*). Another pattern is the citation marker followed by a capitalized noun and a number (e.g. *"[73] Lemma 7"*). This lexical regularity allows us to identify target section specific citations in an automated fashion. Specifically, we search the entirety of our 29 M citation contexts for word sequences, that match either of the part of speech tag patterns `NNP CD IN <citationmarker>` and `<citationmarker> NNP CD`. Doing this, we find 365,299 matches (1.25% of all contexts). This is less then the 2.31% one would expect due to the manual analysis[25] and suggests, that above two patterns are not exhaustive. Nevertheless we can use the identified contexts to further analyze them with respect to their distribution of disciplines.

Table 9 shows the results of this subsequent analysis. Because our data set does not contain equal numbers of citations from each discipline (cf. Fig. 7), we normalize the absolute numbers of pattern occurrences. Rows are then sorted by normalized ratio in decreasing order. Looking at the citing documents (those in which the pattern was found), we see a similar picture to the one in our manual analysis (shown in Table 8). Namely, mathematics with the highest count of target section specific citations by far, and a similar count for computer science and physics, where the latter is slightly lower. Counting by the cited documents (the document in which a specific part is being referenced), the differences decrease a little bit, but mathematics still occurs most frequently by far.

An interesting pattern emerges, when taking an even more detailed look and breaking these citations down by the disciplines on *both* sides of the citation relation. We then can observe the following.

- The most determining factor for target section specific citations seems to be, that a mathematician is writing the document.[†] As with integral and syntactic citations, the writing culture of the field might play a role here.
- The second most determining factor then appears to be, that a mathematical paper is being cited.[‡] Mathematics documents might lend themselves to being cited in this way.
- The third most determining factor is an intra-discipline citation (i.e. the citing document is from that same discipline as the cited). This supports the interpretation of tar-

[25] Because disciplines are not equally represented in the data set, the expected value is not simply the average of values in Table 8 ($\frac{5+17+4+7}{4} \times 300^{-1} = 0.0275$), but a weighted average $(5 \times w_{\text{cs}} + 17 \times w_{\text{math}} + 4 \times w_{\text{phys}} + 7 \times w_{\text{other}}) \times 300^{-1}$, with $\sum w_{\langle \text{discipline} \rangle} = 1$. This gives a value of ≈ 0.0231.

Table 9 Occurrence of target section specific citations by discipline (pairs annotated as follows, †: Mathematics citing document, ‡: Mathematics cited document, X → X: Citing and cited document are from the same discipline)

Discipline	Count	Normalization factor	Normalized ratio (%)
Citing			
Mathematics	298,009	4.66	8.70
CS	9,123	6.31	0.36
Physics	30,593	1.72	0.33
Cited			
Mathematics	313,651	3.15	6.20
CS	12,179	8.50	0.65
Physics	31,087	2.04	0.40
Pairs			
Math† → Math‡	200,859	5.41	6.81
Math† → CS	5,134	92.13	2.96
Math† → Phys	3,114	89.88	1.75
CS → Math‡	3,456	18.82	0.41
Phys → Math‡	3,859	16.49	0.40
CS → CS	2,500	11.38	0.18
Phys → Phys	10,374	2.12	0.14
CS → Phys	50	307.16	0.10
Phys → CS	137	101.40	0.09

get section specific citations as a sign of familiarity with what is being cited (cf. section 'Integral', 'syntactic' and 'target section specific' citations).

Math → Math pairs, where all three of the above factors come into play simultaneously, consequentially show the highest occurrence of target section specific citations by far.

To summarize the results of our analysis of citation flow and citation contexts, we note the following points.

- Publications in mathematics are cited from "outside the field" (e.g. by computer science or physics papers) to a comparatively high degree. Distinguishing citation contexts referring to mathematics publications by discipline might therefore be beneficial in certain applications (e.g. citation-based automated survey generation).
- For most publications, only one or a few citation contexts are available.
- Integral citations appear to be about twice as common in computer science as they are in physics, and again twice as common in mathematics as they are in computer science. Going with Swale's interpretation of the phenomenon, this would mean the focus put on authors in mathematics is higher than in computer science, and higher in computer science than in physics.
- In mathematics, syntactic citations seem to be more common that non-syntactic citations. This is beneficial for reference scope identification (Abu-Jbara and Radev 2012) and any sophisticated approaches based on citation contexts (like context-aware citation recommendation), as citation markers in syntactic citations stand in a grammatical relation to their surrounding words.

– We define target section specific citations as those citations, where a specific section within the cited document is referred to. This type of citation is the most common in mathematics (comparing mathematics, computer science and physics). Through an subsequent analysis of 365k target section specific citations, we find that they are more common in intra-discipline citations than in inter-discipline citations. This supports our assumption that they are an indicator for familiarity with the cited document.

Our five criteria outlined in the beginning, namely *size*, *cleanliness*, *global citation annotations*, *data set interlinkage*, *cross-domain coverage*, in the end made it possible to reach above results. Without sufficient size, our results would be less informative. If our documents contained too much noise, the quality of reference resolution would have deteriorated. Global citation annotations, especially because of their word level precision, make fine grained lexical analyses of citation contexts like the one in section "Automated analysis of target section specific citations" possible. Without interlinking our data set to the MAG, available meta data would have been scarce. While we mainly focused on the scientific discipline information in the MAG, there is much more (authors, venues, etc.) that can be worked with in future analyses. Lastly, if our data set would have only covered a single scientific discipline, an analysis of citation flow, as well as interdisciplinary comparisons of citation context criteria would not have been possible.

Conclusion

Evaluating and applying approaches to research paper-based and citation-based tasks typically requires large, high-quality, citation-annotated, interlinked data sets. In this paper, we proposed a new data set with over one million papers' full texts, 29.2 million annotated citations, and 29.2 million extracted citation contexts (of three sentences each), ready to be used by researchers and practitioners. We provide the data set and the implementation for creating the data set from arXiv source files online for further usage.

For the future, we plan to use the data set for a variety of tasks. Among others, we will develop a citation recommendation system based on all arXiv papers. Furthermore, we plan to perform additional analyses on citations and citation contexts across scientific disciplines, and to use the differences in citing behavior for enhanced citation recommendation.

Acknowledgements Open Access funding provided by Projekt DEAL.

References

Abu-Jbara, A., & Radev, D. (2012). Reference scope identification in citing sentences. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics:*

human language technologies, association for computational linguistics, Stroudsburg, PA, USA (pp. 80–90).

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and polarity of citation: Towards NLP-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, association for computational linguistics, Atlanta, Georgia* (pp. 596–606).

Bast, H., & Korzen, C. (2017). A benchmark and evaluation for text extraction from PDF. In *Proceedings of the 2017 ACM/IEEE joint conference on digital libraries*, JCDL'17 (pp. 99–108).

Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, *17*(4), 305–338. https://doi.org/10.1007/s00799-015-0156-0.

Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M., Lee, D., Powley, B., Radev, D.R., & Tan, Y.F. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the sixth international conference on language resources and evaluation*, LREC'08.

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133.

Caragea, C., Wu, J., Ciobanu, A.M., Williams, K., Ramírez, J.P.F., Chen, H., Wu, Z., & Giles, C.L. (2014). CiteSeer x : A scholarly big dataset. In *Proceedings of the 36th European conference on IR research*, ECIR'14 (pp. 311–322).

Chakraborty, T., & Narayanam, R. (2016). All fingers are not equal: Intensity of references in scientific articles. In *Proceedings of the 2016 conference on empirical methods in natural language processing, EMNLP'16* (pp. 1348–1358).

Chandrasekaran, M.K., Yasunaga, M., Radev, D.R., Freitag, D., & Kan, M. (2019). Overview and results: CL-SciSumm shared task 2019. In *Proceedings of the 4th joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries*, BIRNDL'19, (pp. 153–166).

Chen, J., & Zhuge, H. (2019). Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, *31*(3), e4261.

Duma, D., Klein, E., Liakata, M., Ravenscroft, J., & Clare, A. (2016). Rhetorical classification of anchor text for citation recommendation. *D-Lib Magazine*, *22*, 1.

Ebesu, T., & Fang, Y. (2017). Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, SIGIR'17, (pp. 1093–1096).

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D. J., & Radev, D. R. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the Association for Information Science and Technology*, *59*(1), 51–62. https://doi.org/10.1002/asi.20707.

Färber, M., & Sampath, A. (2019). Determining how citations are used in citation contexts. In *Proceedings of the 23th international conference on theory and practice of digital libraries*, TPDL'19.

Färber, M., Thiemann, A., & Jatowt, A. (2018). A high-quality gold standard for citation-based tasks. In *Proceedings of the 11th international conference on language resources and evaluation*, LREC'18.

Galke, L., Mai, F., Vagliano, I., & Scherp, A. (2018). Multi-modal adversarial autoencoders for recommendations of citations and subject labels. In *Proceedings of the 26th conference on user modeling, adaptation and personalization, ACM, New York, NY, USA, UMAP '18* (pp. 197–205). https://doi.org/10.1145/3209219.3209236.

Ghosh, S., Das, D., & Chakraborty, T. (2016). Determining sentiment in citation text and analyzing its impact on the proposed ranking index. In *Proceedings of the 17th international conference on computational linguistics and intelligent text processing*, CICLing'16 (pp. 292–306).

Gipp, B., Meuschke, N., & Lipinski, M. (2015). CITREC: An evaluation framework for citation-based similarity measures based on TREC genomics and PubMed central. In *Proceedings of the iConference 2015*.

He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, C.L. (2010). Context-aware citation recommendation. In *Proceedings of the 19th international conference on world wide web*, WWW'10, (pp. 421–430).

Huang, W., Wu, Z., Liang, C., Mitra, P., & Giles, C.L. (2015). A neural probabilistic model for context based citation recommendation. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence, AAAI Press, AAAI'15* (pp. 2404–2410).

Huh, S. (2014). Journal article tag suite 1.0: National information standards organization standard of journal extensible markup language. *Science Editing*, *1*(2), 99–104. https://doi.org/10.6087/kcse.2014.1.99.

Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, *20*(3), 341–367. https://doi.org/10.1093/applin/20.3.341.

Lamers, W., Eck, N.J.v., Waltman, L., & Hoos, H. (2018). Patterns in citation context: the case of the field of scientometrics. In *STI 2018 conference proceedings, centre for science and technology studies (CWTS)* (pp 1114–1122).

Liang, L., Rousseau, R., & Zhong, Z. (2013). Non-english journals and papers in physics and chemistry: Bias in citations? *Scientometrics*, *95*(1), 333–350. https://doi.org/10.1007/s11192-012-0828-0.

Liu, F., Hu, G., Tang, L., & Liu, W. (2018). The penalty of containing more non-english articles. *Scientometrics*, *114*(1), 359–366. https://doi.org/10.1007/s11192-017-2577-6.

Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and advanced technology for digital libraries* (pp. 473–474). Berlin: Springer.

Mohammad, S., Dorr, B.J., Egan, M., Awadallah, A.H., Muthukrishnan, P., Qazvinian, V., Radev, D.R., Zajic, D.M. (2009). Using citations to generate surveys of scientific paradigms. In *Proceedings of the 2009 annual conference of the North American chapter of the association for computational linguistics*, NAACL-HLT'09, (pp. 584–592).

Mohapatra, D., Maiti, A., Bhatia, S., & Chakraborty, T. (2019). Go wide, go deep: Quantifying the impact of scientific papers through influence dispersion trees. In *Proceedings of the 19th ACM/IEEE joint conference on digital libraries*, JCDL'19 (pp. 305–314).

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*(1), 86–92.

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2018). Information extraction from scientific articles: A survey. *Scientometrics*, *117*(3), 1931–1990. https://doi.org/10.1007/s11192-018-2921-5.

Prasad, A., Kaur, M., & Kan, M. Y. (2018). Neural ParsCit: A deep learning based reference string parser. *International Journal on Digital Libraries*, *19*, 323–337.

Radev, D. R., Muthukrishnan, P., Qazvinian, V., & Abu-Jbara, A. (2013). The ACL anthology network corpus. *Language Resources and Evaluation*, *47*(4), 919–944.

Reingewertz, Y., & Lutmar, C. (2018). Academic in-group bias: An empirical examination of the link between author and journal affiliation. *Journal of Informetrics*, *12*(1), 74–86. https://doi.org/10.1016/j.joi.2017.11.006.

Roy, D., Ray, K., & Mitra, M. (2016). From a scholarly big dataset to a test collection for bibliographic citation recommendation. *AAAI Workshops*. https://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12635.

Saier, T., & Färber, M. (2019). Bibliometric-enhanced arXiv: A data set for paper-based and citation-based tasks. In *Proceedings of the 8th international workshop on bibliometric-enhanced information retrieval (BIR 2019) co-located with the 41st European conference on information retrieval (ECIR 2019), Cologne, Germany, April 14, 2019*, (pp. 14–26).

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.P., & Wang, K. (2015). An overview of microsoft academic service (MAS) and applications. In *Proceedings of the 24th international conference on world wide web*, WWW'15, (pp. 243–246).

Sugiyama, K., & Kan, M. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. *International Journal on Digital Libraries*, *16*(2), 91–109.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tang, X., Wan, X., & Zhang, X. (2014). Cross-language context-aware citation recommendation in scientific articles. In *Proceedings of the 37th international ACM SIGIR conference on research and development in information retrieval, ACM, SIGIR '14* (pp. 817–826). https://doi.org/10.1145/2600428.2609564.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a) An annotation scheme for citation function. In *Proceedings of the 7th SIGdial workshop on discourse and dialogue, association for computational linguistics, SigDIAL '06* (pp. 80–87).

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b) Automatic classification of citation function. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, EMNLP'06, (pp. 103–110).

Tkaczyk, D., Szostek, P., Fedoryszak, M., Dendek, P. J., & Bolikowski, L. (2015). CERMINE: Automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)*, *18*(4), 317–335.

Tkaczyk, D., Collins, A., Sheridan, P., & Beel, J. (2018). Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. In *Proceedings of*

the 18th ACM/IEEE on joint conference on digital libraries, ACM, New York, NY, USA, JCDL '18 (pp. 99–108). https://doi.org/10.1145/3197026.3197048.

Valenzuela, M., Ha, V., & Etzioni, O. (2015). Identifying meaningful citations. *AAAI Workshops*. https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10185.

Whidby, M., Zajic, D., & Dorr, B. (2011). Citation handling for improved summarization of scientific documents. Tech. rep.