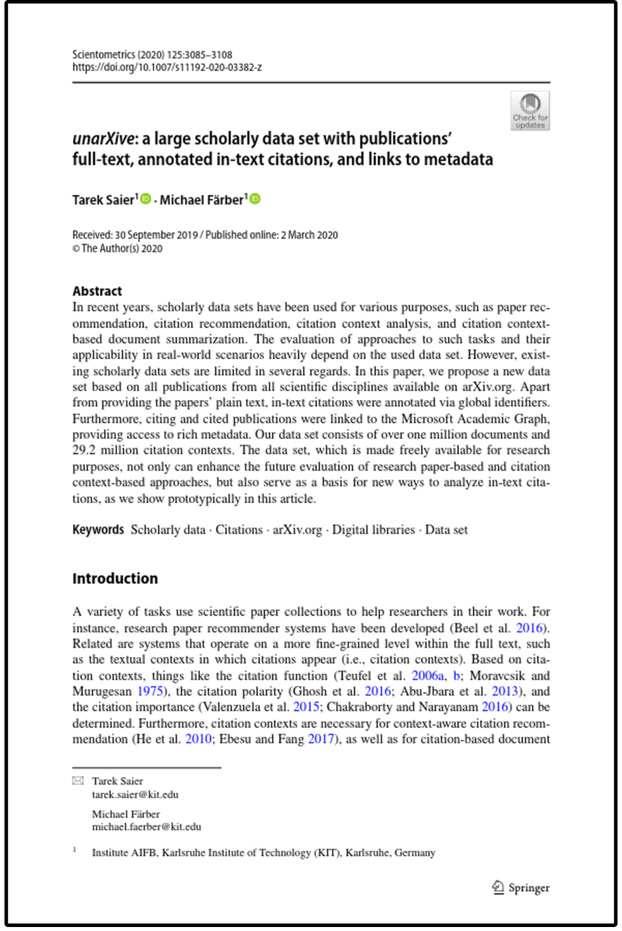
PDF



Scientometrics (2020) 125:3085–3108
https://doi.org/10.1007/s11192-020-03382-z

Check for updates

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 / Published online: 2 March 2020
© The Author(s) 2020

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

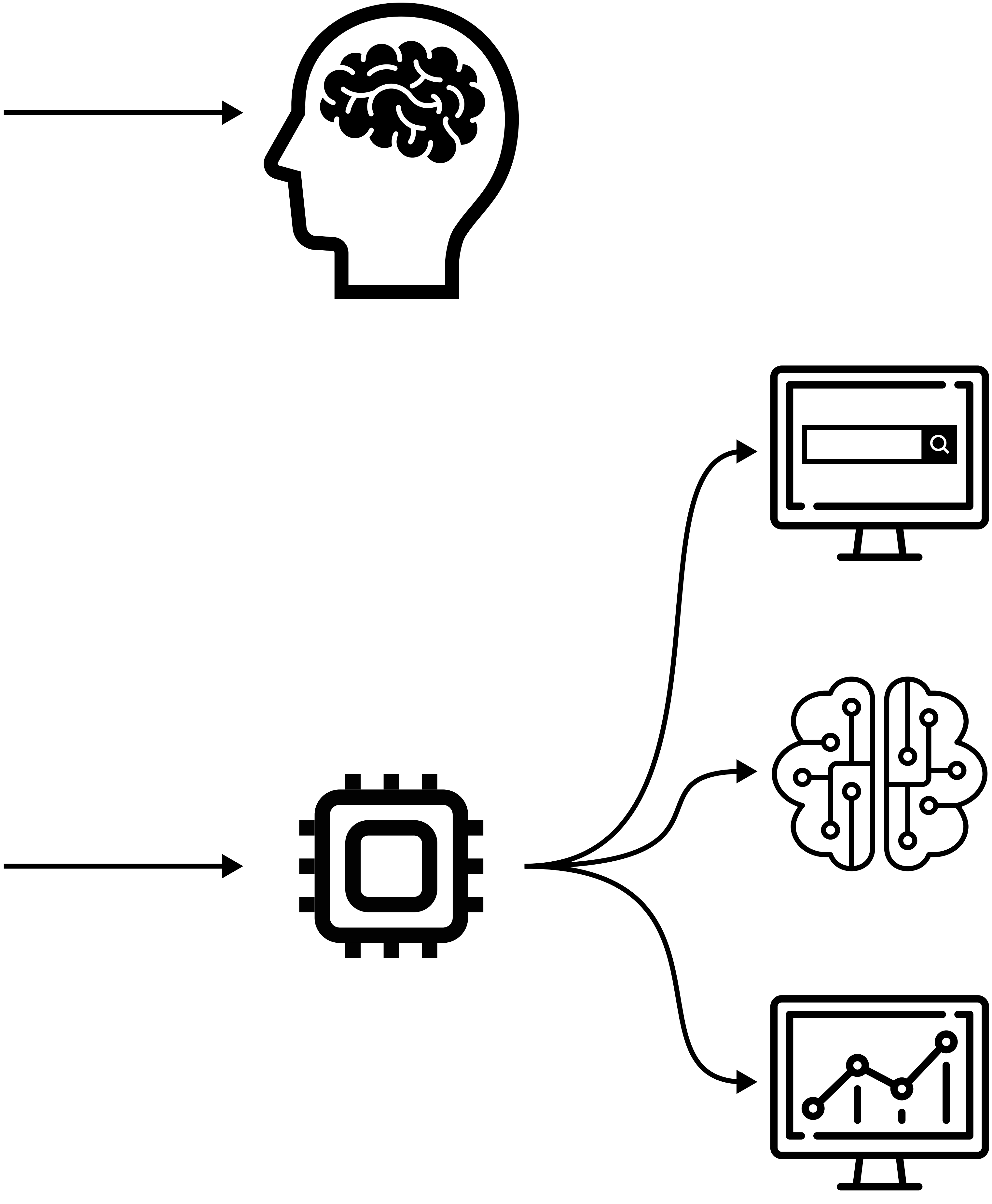
A variety of tasks use scientific paper collections to help researchers in their work. For instance, research paper recommender systems have been developed (Beel et al. 2016). Related are systems that operate on a more fine-grained level within the full text, such as the textual contexts in which citations 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

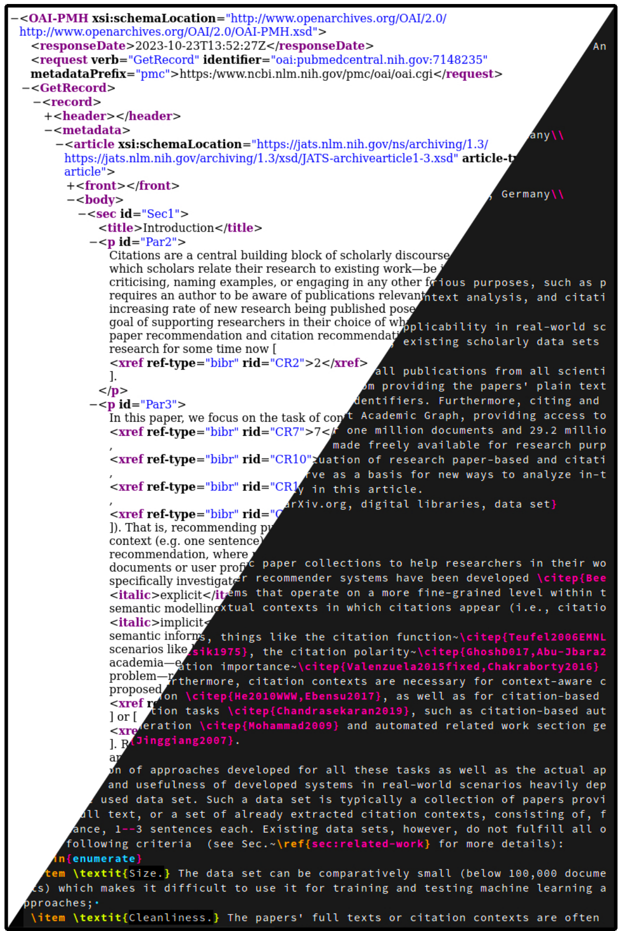
Michael Färber
michael.faerber@kit.edu

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

Springer



XML / LaTeX



```
−<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
 http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2023-10-23T13:52:27Z</responseDate>
   <request verb="GetRecord" identifier="oai:pubmedcentral.nih.gov:7148235"
   metadataPrefix="pmc">https:/www.ncbi.nlm.nih.gov/pmc/oai/oai.cgi</request>
  −<GetRecord>
    −<record>
      +<header></header>
      −<metadata>
        −<article xsi:schemaLocation="https://jats.nlm.nih.gov/ns/archiving/1.3/
          https://jats.nlm.nih.gov/archiving/1.3/xsd/JATS-archivearticle1-3.xsd" article-t
          article">
          +<front></front>
          −<body>
            −<sec id="Sec1">
               <title>Introduction</title>
              −<p id="Par2">
                 Citations are a central building block of scholarly discourse
                 which scholars relate their research to existing work—be i
                 criticising, naming examples, or engaging in any other f
                 requires an author to be aware of publications relevant
                 increasing rate of new research being published pose
                 goal of supporting researchers in their choice of wh
                 paper recommendation and citation recommendati
                 research for some time now [
                 <xref ref-type="bibr" rid="CR2">2</xref>
                 ].
               </p>
              −<p id="Par3">
                 In this paper, we focus on the task of con
                 <xref ref-type="bibr" rid="CR7">7</
                 ,
                 <xref ref-type="bibr" rid="CR10"
                 ,
                 <xref ref-type="bibr" rid="CR1
                 ,
                 <xref ref-type="bibr" rid="C
                 ]). That is, recommending pu
                 context (e.g. one sentence)
                 recommendation, where
                 documents or user profi
                 specifically investigate
                 <italic>explicit</it
                 semantic modellin
                 <italic>implicit<
                 semantic inform
                 scenarios like
                 academia—e
                 problem—r
                 proposed
                 <xref r
                 ] or [
                 <xre
                 ]. R
                 ar
```

```
                                                                 An
                                                           any\\
                                                       , Germany\\
                                     rious purposes, such as p
                                    ntext analysis, and citati
                                   pplicability in real-world sc
                                   , existing scholarly data sets
                                 all publications from all scienti
                                om providing the papers' plain text
                               dentifiers. Furthermore, citing and
                              t Academic Graph, providing access to
                             r one million documents and 29.2 millio
                             made freely available for research purp
                            uation of research paper-based and citati
                           rve as a basis for new ways to analyze in-t
                          y in this article.
                         arXiv.org, digital libraries, data set}
                     c paper collections to help researchers in their wo
                    r recommender systems have been developed \citep{Bee
                   ems that operate on a more fine-grained level within t
                  xtual contexts in which citations appear (i.e., citatio
                 s, things like the citation function~\citep{Teufel2006EMNL
                sik1975}, the citation polarity~\citep{GhoshD017,Abu-Jbara2
               ation importance~\citep{Valenzuela2015fixed,Chakraborty2016}
              rthermore, citation contexts are necessary for context-aware c
             on \citep{He2010WWW,Ebensu2017}, as well as for citation-based
            tion tasks \citep{Chandrasekaran2019}, such as citation-based aut
           eration \citep{Mohammad2009} and automated related work section ge
          {Jinggiang2007}.
         on of approaches developed for all these tasks as well as the actual ap
         and usefulness of developed systems in real-world scenarios heavily dep
        used data set. Such a data set is typically a collection of papers provi
       ull text, or a set of already extracted citation contexts, consisting of, f
      ance, 1--3 sentences each. Existing data sets, however, do not fulfill all o
      following criteria  (see Sec.~\ref{sec:related-work} for more details):
    in{enumerate}
   tem \textit{Size.} The data set can be comparatively small (below 100,000 docume
  ts) which makes it difficult to use it for training and testing machine learning a
 pproaches;
 \item \textit{Cleanliness.} The papers' full texts or citation contexts are often
```