





Structured References from PDF Articles: Assessing the Tools for Bibliographic Reference Extraction and Parsing

Alessia Cioffi¹  and Silvio Peroni^{2,3} 

- ¹ Digital Humanities and Digital Knowledge, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
alessia.cioffi@studio.unibo.it
- ² Research Centre for Open Scholarly Metadata, Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
silvio.peroni@unibo.it
- ³ Digital Humanities Advanced Research Centre (DH.arc), Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

Abstract. Many solutions have been provided to extract bibliographic references from PDF papers. Machine learning, rule-based and regular expressions approaches were among the most used methods adopted in tools for addressing this task. This work aims to identify and evaluate all and only the tools which, given a full-text paper in PDF format, can recognise, extract and parse bibliographic references. We identified seven tools: Anystyle, Cermine, ExCite, Grobid, Pdfssa4met, Scholarcy and Science Parse. We compared and evaluated them against a corpus of 56 PDF articles published in 27 subject areas. Indeed, Anystyle obtained the best overall score, followed by Cermine. However, in some subject areas, other tools had better results for specific tasks.

Keywords: References extraction · References parsing · Structured citation data

1 Introduction

In past decades, the academic publishing world has needed to face an exponential increase in the volume of scientific literature materials [13, 29]. The necessity to handle such a vast amount of information has been one of the drivers of the digitalisation of literature materials. The conversion of academic knowledge to structured and machine-readable formats revealed positive effects also in the searchability and availability of such information, thanks to services like search engines [19]. At the same time, the structured format allowed us to valorise the citation graph connecting the scientific literature [7]. Also, in the past 50 years, bibliographic references have assumed a more prominent role in the scientific community, not only for tracking evolution in science but also for measuring impact [14].

In the past five years, the Initiative for Open Citations (I4OC, <https://i4oc.org>) has emphasised the importance of making citation data public. One of the main challenges

to address for reaching this goal concerns extracting them from unstructured documents, like PDFs, and converting them into structured data in specific formats (e.g. JSON, XML, RDF). However, such extraction is made even more complex by the variety of (either standard or ad hoc) reference styles [15].

In the past, several tools have been proposed to address this task. Our work aims to analyse the current availability of these tools to identify which outperforms the others in extracting and parsing bibliographic references of academic papers.

The rest of the paper is structured as follows. In Sect. 2, we introduce the methodology adopted for identifying relevant tools and analyse their performance against a gold standard. The outcomes of the tools are shown in Sect. 3 and are discussed in more detail in Sect. 4. In Sect. 5, we introduce some of the essential related works in reference extraction approaches and tools. Finally, Sect. 6 concludes the work by sketching out some future developments.

2 Materials and Methods

We devised a methodology for the identification and evaluation of the reference extraction tools, which is based on four steps: (a) systematic literature review, (b) creation of a dataset, (c) creation of translation scripts, and (d) evaluation scripts. Following [31], a specific procedure was implemented and formalised in a protocol fully described in [5] – which is not reported entirely here for page constraints. Such a protocol is based on a citation-based search strategy [30] and uses seed papers for starting the search process [18]. In the first step (a), we decided to consider only papers written in English and dated after 2005. Once relevant articles were chosen in the literature, the focus moved to identify the reference extraction tools described in such documents. We decided to consider, in the analysis, only the tools that can parse full-text PDF papers, retrieve singularly tagged references, retrieve the metadata of each reference, and be either a standalone application or a programming language library, including APIs. At the end of this step, we have identified the following tools: Anystyle (<https://github.com/inukshuk/anystyle-cli>), CERMINE [26], EXCITE [17], GROBID [20], PDFSSA4MET (<https://github.com/eliask/pdfssa4met>), Scholarcy [8], and Science Parse (<https://github.com/allenai/science-parse>).

The next step (b) concerned preparing the data to use to test the tools identified. An initial dataset of papers in PDF format was selected to be processed by the reference extraction tools to obtain these data. This dataset included academic papers from different research fields from a corpus of selected articles used in a complementary study [24]. The dataset comprised 2,538 bibliographic references referring to almost 1,000 different journals, extracted from two articles for each one of the following 27 subject areas: Agricultural and Biological Sciences (AGR-BIO-SCI), Arts and Humanities (ART-HUM), Biochemistry, Genetics and Molecular Biology (BIO-GEN-MOL), Business, Management and Accounting (BUS-MAN-ACC), Chemical Engineering (CHE-ENG), Chemistry (CHEM), Computer Science (COM-SCI), Decision Sciences (DEC-SCI), Dentistry (DEN), Earth and Planetary Sciences (EAR-PLA-SCI), Economics, Econometrics and Finance (ECO-ECO-FIN), Energy (ENE), Engineering (ENG), Environmental Science

(ENV-SCI), Health Professions (HEA-PRO), Immunology and Microbiology (IMM-MIC), Materials Science (MAT-SCI), Mathematics (MAT), Medicine (MED), Multi-disciplinary (MUL), Neuroscience (NEU), Nursing (NUR), Pharmacology, Toxicology and Pharmaceutics (PHA-TOX-PHA), Physics and Astronomy (PHY-AST), Psychology (PSY), Social Sciences (SOC-SCI), Veterinary (VET). These were complemented with additional two articles having bibliographic references not introduced in a ‘References’ or ‘Literature’ section (Z-NOTES-TEST). We created a gold standard for comparing the outcomes of the reference extraction tools from these papers. We used the common metadata defining bibliographic references according to the analysis run in [24] as a baseline to understand which metadata must be identified and marked in each bibliographic reference depending on the type of the cited object.

The following step (c) consisted of translating the output of the reference extraction tools into the same format (TEI was chosen) to enable automatic comparison of such output with the gold standard. Finally, we evaluated (d) the tools using precision, recall and f-score, according to the following dimensions (based on prior studies [12, 27]):

1. *Correctly identified references.* The software’s ability to distinguish each reference from the surrounding text and other references. The aim is to determine how many references are correctly identified by each parser.
2. *Correctly identified fields per reference.* The number of correctly tagged metadata, independently from content correctness. This analysis allows us to check the tools’ quality of the markers’ usage.
3. *Correctly identified contents per reference.* How many parts of the bibliographic reference have been correctly parsed and tagged for verifying if the text inside a correctly identified metadata is correct.

The software and all the data used for the experiment are available in [3] and [4].

3 Results

The overall results of the tools’ assessment, introduced in Table 1, showed that Anystyle had the best performance. Nonetheless, it is possible to see a different distribution of the values between references, metadata and contents. As expected, the lowest f-score was retrieved in the correct identification of references since it was derived from the correct identification of the metadata elements and their content. Cermin showed its lowest f-score in the references dimension and its highest f-score in the metadata element identification. Overall, the dimension related to metadata contents showed that, even if the metadata element was correctly identified, the content it contained was prone to parsing errors.

The results per subject area, summarised in Fig. 1, differed slightly from the overall ones. Indeed, Anystyle showed coherent results, with all f-values above 0.5 and the highest value registered at 0.97 (BUS-MAN-ACC in Fig. 1). Another noticeable aspect is the high quality of the identification of references in the set of files which included bibliographic references in a section not labelled as “References” or “Literature” (Z-NOTES-TEST), whose p-value lay above 0.85.

Table 1. Precision (P), recall (R) and f-score (F1) of each dimension analysed per tool.

Tools	References			Metadata			Content		
	P	R	F1	P	R	F1	P	R	F1
Anystyle	0.81	0.74	0.77	0.93	0.97	0.95	0.87	0.91	0.89
Cerminc	0.75	0.67	0.71	0.94	0.94	0.94	0.86	0.87	0.86
ExCite	0.59	0.53	0.56	0.93	0.92	0.92	0.79	0.79	0.79
Grobid	0.54	0.55	0.54	0.86	0.97	0.91	0.81	0.92	0.86
Pdfssa4met	0.01	0.14	0.07	0.01	0.29	0.14	0.01	0.19	0.09
Scholarcy	0.62	0.78	0.69	0.96	0.70	0.81	0.90	0.65	0.75
Science Parse	0.43	0.32	0.37	1.00	0.55	0.71	0.94	0.51	0.66

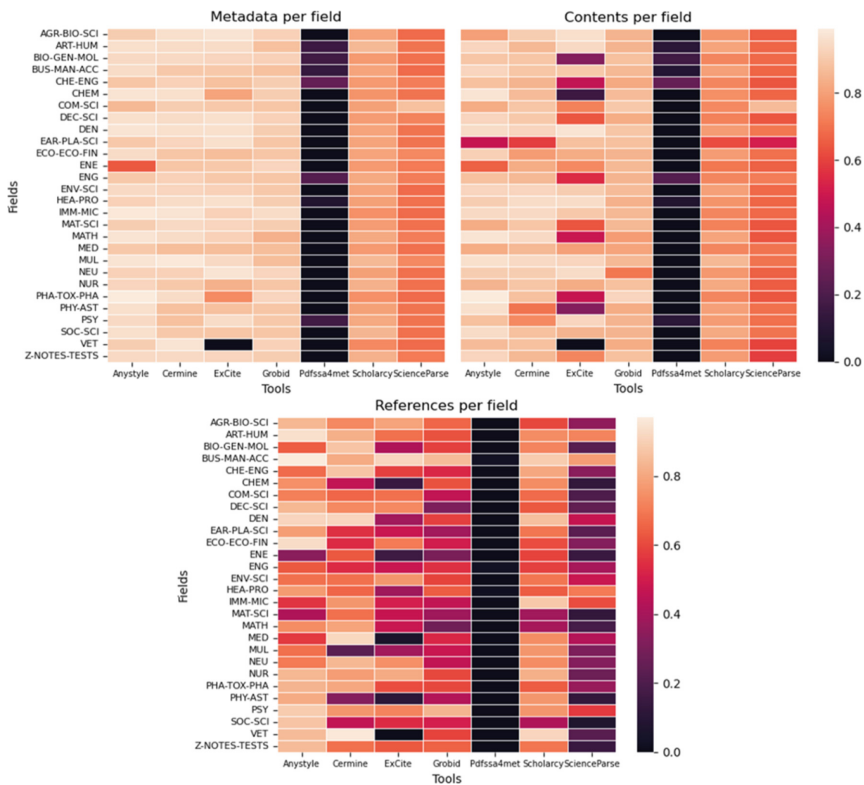


Fig. 1. Comparison of the f-scores per subject areas (i.e. fields) in references identification.

Also, Cerminc showed a high precision in reference identification, with a maximum score of 0.96 and a minimum of 0.23. The values were distributed among the fields so that, while only a few fields presented high values above 0.9, many of the fields were close to slightly lower values ranging between 0.6 and 0.8.

ExCite showed high f-scores for reference identification (e.g. 0.91 in BUS-MAN-ACC), with related high f-scores in metadata and content identification (e.g. 0.98

and 0.97, respectively, for the same subject area). However, it could not identify any bibliographic reference in the articles in VET.

The f-scores gathered using GROBID varied a lot in assessing reference identification (from 0.28 to 0.85), but they showed a smaller range in identifying contents (from 0.71 to 0.93). Pdfssa4met, instead, was the tool showing the worst performances. It was able to identify a few references (and related metadata) only in seven subject areas and showed a very low precision (from 0.01 to 0.03).

Scholarcy's f-scores highlighted the excellent performances of the tool in the main part of the subject areas, where the f-scores for the identification of references, metadata element and related content were greater than 0.58, 0.87 and 0.75, respectively. Finally, Science Parse had 0.78 as the maximum f-score in the task of reference identification (in the BUS-MAN-ACC subject area). It is worth mentioning that the precision was 1.0 in all the fields. This was not unexpected since this tool could identify only four metadata elements in each reference (i.e. author, title, source and year), thus reducing the chances of mismatching different elements.

4 Discussion

The comparison between tools' output and the gold standard showed a complex scenario in which a tool, Anystyle, outperformed the others. Indeed, Anystyle obtained the best score in all three dimensions of the analysis, i.e. references, metadata and contents, followed by Cermin. The remaining tools showed good performances on average, except Pdfssa4met.

It is worth mentioning that other factors that affected the reference extraction by the tools were the citation practice of particular subject areas since it affected the results mainly due to the different writing and collecting references practices. Indeed, reference identification was very effective in some subject areas, but other areas (e.g. ENE) showed low performance in all the tools. Thus, it came out that the tools' performances are affected by the practices in the subject areas and that none of the tools was good per se in all the subject areas.

This work presents three major limitations. First, the input dataset was small, even if appropriate to run initial experiments on the topic. Indeed, even if providing a vast number of research fields, each subject area included only two papers, enough to provide a preliminary insight rather than a definitive view on the topic. Second, the tools have been used off-the-shelf, without any training. For the CRF-based tools, this lack of training could have resulted in a loss in performance for some of the tools [28]. Finally, we adopted the Levenshtein distance as a unique metric to compute the similarity of the metadata content in the bibliographic references. Nonetheless, other works have identified other measures, e.g. the soft TF-IDF [6], to outperform the Levenshtein distance in measuring the similarity between two names in text retrieval tasks.

5 Related Works

Apart from the tools identified and used in our analysis, we took notes about other theoretical approaches and workflows presented in other articles when we identified

the tools to use in our study. This section presents some of the most important ones, organised in three categories.

Single Reference Parsing. This category of tools represents a set of tools which can parse a single reference and returns the metadata it is composed of in a structured format. The tools can be different depending on the approach they are based on, the input data they accept, the focus on different types of citation, e.g. academic or generic references, or the ability to extract a different number of metadata from the reference strings. Some of these tools are based on machine learning techniques, e.g. [33], while others use Hidden Markov Model [9, 32, 22], rule-based methods [25] and frame-based approaches [10] to address the same tasks.

Parsers for Reference Lists. This is a category of tools that extract and parse references from files in different formats, but not from full-text pdf files. Indeed, in most cases, they can, given a text file with a list of references (one line per reference), extract single references, parse them and return the metadata of each reference, such as Neural Parscit (<https://github.com/WING-NUS/Neural-ParsCit>).

Frameworks for Parsing Bibliographic References in PDF Full Text. In [23], the authors describe a machine-learning-based framework that outperforms the results obtained on the same input dataset by an HMM-based method. Similarly, in [26], the authors explore a composed tool based on simple HMM and rules thought to be easily modifiable by the user. Other solutions are based on rules, e.g. [1, 11, 16], ontologies [21], or deep pully convolutional networks [2].

6 Conclusions

This work aimed to retrieve from the available literature all the tools able to extract the bibliographic references from full-text PDF papers and evaluate them. Seven tools have been selected: Anystyle, Cermine, ExCite, Grobid, Pdfssa4met, Scholarcy and Science Parse. Three dimensions have been analysed for each: the correctly extracted metadata, the related correctly extracted contents, and the correctly extracted references.

Anystyle outperformed the others in all the three dimensions considered in the analysis. Nonetheless, the results for the analysis per subject area showed that, in some cases, Anystyle was outperformed by other tools. Thus, while Anystyle is the best tool for bibliographic reference extraction and parsing, cooperation between the tools based on the specific subtasks may be relevant to obtaining the best possible results.

In future developments, extending the current corpus of input PDF documents could be appropriate to consolidate the results obtained in this research.

Acknowledgements. The work of Silvio Peroni has been partially funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 101017452 (OpenAIRE-Nexus).

References

1. Azimjonov, J., Alikhanov, J.: Rule based metadata extraction framework from academic articles. *arXiv:1807.09009* [Cs] (2018)
2. Bhardwaj, A., Mercier, D., Dengel, A., Ahmed, S.: DeepBIBX: deep learning for image based bibliographic data extraction. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, E.-S. M. (eds.) *Neural Information Processing*, pp. 286–293. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-70096-0_30
3. Cioffi, A.: Code for converting different formats to TEI XML and evaluation of the results. Zenodo (2022). <https://doi.org/10.5281/zenodo.6182128>
4. Cioffi, A.: Data for testing and evaluating references extraction and parsing tools. Zenodo (2022). <https://doi.org/10.5281/zenodo.6182066>
5. Cioffi, A.: Systematic literature review about software for references extraction. protocols.io (2022). <https://doi.org/10.17504/protocols.io.buz9nx96>
6. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: *IIWEB 2003: Proceedings of the 2003 International Conference on Information Integration on the Web (2003)*. <https://doi.org/10.5555/3104278.3104293>
7. Fortunato, S., et al.: Science of science. *Science* **359**(6379), aao0185 (2018). <https://doi.org/10.1126/science.aao0185>
8. Gooch, P.: How Scholarcy contributes to and makes use of open citations. Scholarcy (2021). <https://www.scholarcy.com/how-scholarcy-contributes-to-and-makes-use-of-opencitations/>
9. Hetzner, E.: A simple method for citation metadata extraction using hidden Markov models. In: *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL 2008*, p. 280. Pittsburgh PA, PA, USA: ACM Press (2008)
10. Hsieh, Y.L., et al.: A frame-based approach for reference metadata extraction. In: Cheng, S.M., Day, M.Y. (eds.) *Technologies and Applications of Artificial Intelligence*. LNCS, vol. 8916, pp. 154–163. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13987-6_15
11. Huynh, T., Hoang, K.: GATE framework based metadata extraction from scientific papers. In: *2010 International Conference on Education and Management Technology*, pp. 188–191. Cairo, Egypt. IEEE (2010). <https://doi.org/10.1109/ICEMT.2010.5657675>
12. Indrawati, A., Yoganingrum, A., Yuwono, P.: Evaluating the quality of the indonesian scientific journal references using ParsCit, CERMINE and GROBID. *Lib. Philos. Pract.* (2019)
13. Khabsa, M., Giles, C.L.: The number of scholarly documents on the public web. *PLoS ONE* **9**(5), e93949 (2014). <https://doi.org/10.1371/journal.pone.0093949>
14. Kim, K., Chung, Y.: Overview of Journal Metrics. *Sci. Editing* **5**(1), 16–20 (2018). <https://doi.org/10.6087/kcse.112>
15. King, D., Jérôme, D., Van Allen, M., Shepherd, P., Bollen, J.: Tools and metrics: keynote speech. *Inf. Serv. Use* **28**(3–4), 215–28 (2009). <https://doi.org/10.3233/ISU-2008-0579>
16. Kluegl, P., Hotho, A., Puppe, F.: Local adaptive extraction of references. In: Dillmann, R., Beyerer, J., Hanebeck, U.D., Schultz, T. (eds.) *KI 2010*. LNCS (LNAI), vol. 6359, pp. 40–47. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16111-7_4
17. Körner, M., Ghavimi, B., Mayr, P., Hartmann, H., Staab, S.: Evaluating reference string extraction using line-based conditional random fields: a case study with German language publications. In: Kirikova, M., et al. (eds.) *ADBIS 2017*. CCIS, vol. 767, pp. 137–145. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67162-8_15
18. Lecy, J.D., Kate, E.: Beatty: representative literature reviews using constrained snowball sampling and citation network analysis. *SSRN Electron. J.* (2012) <https://doi.org/10.2139/ssrn.1992601>
19. Levene, M.: *An Introduction to Search Engines and Web Navigation*, 2nd edn. John Wiley, Hoboken (2010)

20. Lopez, P.: GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonias, G. (eds.) *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, vol. 5714, pp. 473–474. Springer, Berlin (2009). https://doi.org/10.1007/978-3-642-04346-8_62
21. Ning, X., Jin, H., Wu, H.: SemreX: towards large-scale literature information retrieval and browsing with semantic association. In: *2006 IEEE International Conference on E-Business Engineering (ICEBE 2006)*, pp. 602–609. Shanghai, China. IEEE (2006). <https://doi.org/10.1109/ICEBE.2006.87>
22. Ojokoh, B., Zhang, M., Tang, J.: A Trigram hidden Markov model for metadata extraction from heterogeneous references. *Inf. Sci.* **181**(9), 1538–1551 (2011). <https://doi.org/10.1016/j.ins.2011.01.014>
23. Peng, F., Andrew M.: Accurate information extraction from research papers using conditional random fields. In: *NAACL* (2004)
24. Santos, E.A.D., Peroni, S., Mucheroni, M.L.: The way we cite: common metadata used across disciplines for defining bibliographic references. In: *Proceedings of the 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022)*. arXiv.org (2022, to appear). <https://doi.org/10.48550/arXiv.2202.08469>
25. Suryawati, E., Widyantoro, D.H.: Combination of heuristic, rule-based and machine learning for bibliography extraction. In: *2017 5th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, pp. 276–81, Bandung. IEEE (2017). <https://doi.org/10.1109/ICICI-BME.2017.8537772>
26. Tkaczyk, D., Szostek, P., Dendek, P.J., Fedoryszak, M., Bolikowski, L.: CERMINE -- automatic extraction of metadata and references from scientific literature. In: *2014 11th IAPR International Workshop on Document Analysis Systems*, pp. 217–21. IEEE (2014). <https://doi.org/10.1109/DAS.2014.63>
27. Tkaczyk, D., Collins, A., Sheridan, P., Beel, J.: Evaluation and comparison of open source bibliographic reference parsers: a business use case. [arXiv:1802.01168](https://arxiv.org/abs/1802.01168) (2018)
28. Tkaczyk, D., Collins, A., Sheridan, P., Beel, J.: 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*, pp. 99–108. Fort Worth Texas USA. ACM (2018)
29. Van Noorden, R.: Global scientific output doubles every nine years. *nature news blog* (2014). <http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html>
30. Wohlin, C.: Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE 2014* (2014)
31. Xiao, Y., Watson, M.: Guidance on conducting a systematic literature review. *J. Plan. Educ. Res.* **39**(1), 93–112 (2019)
32. Yin, P., Zhang, M., Deng, Z., Yang, D.: Metadata extraction from bibliographies using bigram HMM. In: Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E., Lim, E.-P. (eds.) *ICADL 2004*. LNCS, vol. 3334, pp. 310–319. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30544-6_33
33. Zhang, X., Zou, J., Le, D.X., Thoma, G.R.: A structural SVM approach for reference parsing. *BMC Bioinform.* **12**(S3), S7 (2011). <https://doi.org/10.1186/1471-2105-12-S3-S7>