ALMA MATER STUDIORUM - UNIVERSITA' DI BOLOGNA

Second Cycle Degree Program in

Digital Humanities and Digital Knowledge

Dissertation Title

From PDF to structured references: A comparative study on tools for bibliographic references extraction and parsing

Final dissertation in

Computational Thinking and Programming

Supervisor: Silvio Peroni

Co-supervisor: Fabio Vitali

Presented by: Alessia Cioffi (950430)

Session
III

Academic Year
2021-2022

Table of Contents

Table of Figures

List of Tables

Abstract

In the publishing world an active role is played by bibliographic references. Normally, references are published in PDF format, which represents an issue in the perspective of a digital publishing environment. Many solutions have been provided to extract bibliographic references from papers in PDF format. Machine learning, rule-based and regular expressions are the most widespread methods used for carrying out this task. These methods have been implemented in different ways either in tools or in frameworks.

The aim of this work is to identify all, and only, the tools which, given a full text paper in PDF format, are able to identify, extract and parse bibliographic references. The methods they are based on don't influence the tools selection. The first phase of this thesis is the literature review. From this step, seven tools are identified: Anystyle (client, locally installed), Cermine (locally installed), ExCite (online tool), GROBID (locally installed), Pdfssa4met (locally installed), Scholarcy (online API) and Science Parse (locally installed). In a second moment, these tools are compared and evaluated in different research fields, providing interesting results. Indeed, Anystyle obtains the best overall score, followed by Cermine. However, in some of the subtasks investigated alongside the overall results, other tools resulted to have a better performance in specific tasks. Thus, in this variegated scenario, different solutions can be adopted on the basis on the user's requirements.

1. Introduction

The aim of this thesis is to provide a comparative selection of all the available working tools designed to extract and parse bibliographic references from scholarly articles. There are various reasons behind the development of this work. On the one hand, identifying the relevant tools to carry out the references extraction and parsing tasks is significant in the publication field in order to extract information from PDF papers and make them publicly available. On the other hand, there is the necessity to identify the working tools among the enormous mass of scientific literature on this research topic.

Indeed, in recent decades the academic publishing world needed to face the contemporaneous arising of different issues regarding data management. Initially, it has been registered an exponential increase in the volume of scientific literature materials (Khabsa and Giles 2014; Van Noorden 2014). The consequent necessity to handle such a huge amount of information, has represented a first driver of growth towards the digitalization of literature materials. At the same time, the conversion of academic information to machine readable formats revealed positive effects not only in the information management, but also in the searchability and availability of such information. Indeed, through the presence of services like search engines, there has been an increase in the value of the interconnections between the literature itself (Levene 2010).

Academic publications are internally connected through different typologies of network, e.g. co-authoring or venue networks (Fortunato et al. 2018). One of the most affirmed typologies of literature networking is connected to the publications' references. Indeed, in this perspective, single publications are connected to each other through the references system, either by citing or being cited. Throughout the years the references have also assumed a different and more relevant role in the scientific community. References have become a mean to give prominence to the authors through the analysis of the metrics related to the citations (Kim and Chung 2018). Therefore, having access to the widest possible number of references allows, on the one hand, a better performance on the literature analysis and on the services provided to the final users, and, at the same time, to recognize the values of authors, journals and publishers. Finally, a phenomenon which has been introduced in the last years that is having a relevant impact in the academic world is the Open Science movement. In the optic of this phenomenon, literature information and, in particular, references should be openly available and accessible to everyone. Among other initiatives for Open Science, The Initiative for Open Citations

(I4OC)[1] is emphasizing the necessity of making citation data public on the web in order to enhance the quality of the research availability.

In this scenario a prominent role is played by the format in which data is carried. The Portable Document Format (PDF) is the digital file format in which academic papers, and their relative information, have been mostly published through the years. One of the main issues connected to the conversion of information into the digital environment is related to the necessity of working with structured data. Structured data is data organized in predefined structures, either tabular or graph structured. This kind of data is fundamental in a digital environment due to the fact that structured data is machine readable. Thus, information stored in this format is easier to manage and search (Blumberg and Shaku 2003). Nonetheless, the main format in which academic works are published, PDF, even though it is a digital format, is nonetheless an unstructured type of data. As a consequence, the single items it is composed of, e.g. the references, are not independently manageable. Thus, it becomes crucial, in the perspective of a conversion to a digital environment, the translation of the information carried by unstructured texts into structured data (Rusu et al. 2013). Hence, in this context, the relevance of extracting references from unstructured documents becomes a challenging task. The first issue, in this context, is to find a way to extract information from a type of data not provided with a specific structure. Because of this, different approaches have been proposed to carry out this task during the years, including the use of regular expressions, machine learning or rules. Even if some of them seem to work better than the others, there is not a broad agreement on this matter. The second issue is related to high variance of references, on the basis of the citation style, e.g. APA and Chicago, and the number of fields required, e.g. the article title is not a metadata required by all journals, but rather it is up to the single journal to include it or not (Santos, Peroni and Mucheroni 2022). The different solutions proposed to solve the tasks of extracting bibliographic references from PDF papers, parsing the single references and returning them in a structured format are either processes or tools which must take into account both the aforementioned problems.

Recently, some publishers and editors have started to invest in the publication of the metadata of the published sources, alongside the resources themselves in PDF format. The reason behind this investment can be traced back to the increased relevance of citation metrics. Indeed, this is a source of prestige and economic income for these agents, since they are used to decide, for instance, how to distribute the available institutional funds (Herzog, Hook and Adie 2018). Thus, investing in these operations can provide concrete benefits to publishers. At the same time, smaller publishers may not have enough financial support to carry out this task by their own means, in particular since making

[1] https://i4oc.org/

use of a private tool to extract references requires extra costs with respect to the ones already included in the basic needs of a publisher (King et al. 2009). For these companies, the possibility to identify an open-source tool able to automatically extract and parse references from PDF files could be an optimal solution to this issue. At the same time the literature is full of resolutions developed in order to solve this problem and, in this wideness of options, it may be difficult to identify a good solution, or the best on the basis of the user's needs. In this context, a hopeful result could be to identify a unique open-source tool able to extract the references from a full-text PDF paper and return them in a desired structured output. In this way, the tool under observation is a tool which is able to identify the bibliographic section of a paper and the references which it is composed of, parse these references and therefore identifying the single information which compose it and once identified, returning them in structured form (XML, JSON, BibTeX etc.).

Starting from this principle, this research is aimed at verifying which, among the existing tools in this field, are relevant and whether there are some tools particularly good at carrying out this task. In particular, the references extracting and parsing tools, on which this analysis focuses, given a full-text PDF as input, must be able to extract bibliographic references and return them in a structured form. In order to do this, all the available and usable tools, able to carry out this task, are compared and, on the basis of specified parameters, evaluated. The aim of this process is to analyse the features of the identified tools in order to verify whether they are positively or negatively affecting the references metadata quality. At the end of the analysis the parsers are compared and the results with the annexed discussion will be provided.

To define the research questions guiding this work, it is assumed that the object of the research is the analysis and comparison of tools able to extract and parse references. The objective of these questions is to state the terms in which the tools should be evaluated in order to get some defined results. The research questions can be formulated as follows:

RQ1. Which tools are able to extract bibliographic references from an input full-text PDF paper and return them in a structured format?

RQ2. Is there a tool which outperforms the others in extracting and parsing bibliographic references from academic papers selected from different research fields, e.g. the humanities or computer science? Do some tools have a good or bad performance only in some specific research areas?

RQ3. Is there a tool, or a class of tools, that has a better performance than the others only under specific conditions, e.g. paper layout? Or, vice versa, is there a layout or other features which are particularly difficult to analyse for almost all the tools?

The following work has been organized in order to answer these three questions. The structure of this thesis follows the steps performed to carry out this comparative work. In the next chapter it is presented the literature review, together with the methodology followed in order to carry it out and the results obtained by it. The literature review is the first step performed in order to develop this thesis work. Through it, it is possible to analyse the existing literature and retrieve the tools investigated and compared in this research. The third chapter reports the methodology followed to achieve the final evaluation, the data produced through the application of the methodology and the results obtained by means of the data analysis. Then, the results are discussed in the fourth chapter, with respect to the original research questions. Finally, the conclusions are presented in the fifth chapter, together with the limits of the work and the future developments.

2. Literature Review

2.1 Methodology for Software Identification

A systematic literature review is the first step necessary to retrieve the references extracting tools for this research. The role of a systematic review in the research process is to provide a trustable "means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest" (Kitchenham 2004). Indeed, through a systematic literature review it is possible to navigate the network of publications inherent a specific topic and, on the basis of previously identified and formalized parameters, identify the relevant literature with regards to the stated topic. In the context of the current work, the scope of the systematic literature review is to navigate the literature network and retrieve the papers describing or comparing tools relevant with respect to the topic of bibliographic references extraction and parsing. This process has been performed on the basis of a set of parameters regarding the desired features that the tools must present to be selected. In order to carry out a systematic literature review, one work in particular has been followed, 'Guidance on Conducting a Systematic Literature Review' (Xiao and Watson 2019). On the basis of this work, a specific procedure has been implemented and formalized. First of all, it has been stated the research problem for the review: *which tools are actually able to extract and parse the bibliographic references from full-texts PDF papers?* Then, starting from the stated question, it has been formulated a procedure to identify a solution to it, following the steps described in the work by (Xiao and Watson 2019). The body of the review process, graphically represented in *Figure 2*, is structured into three main steps:

1. Retrieving all the potential papers from the literature tree, following specific research process;
2. Identifying the relevant papers among the ones collected, on the basis of previously defined relevance parameters.
3. Extracting and reporting the data contained in the accepted papers, identifying the valid tools.

All these mentioned steps, followed in order to carry this research out, have been formally listed in a final protocol (Cioffi 2022c). The aim of creating a protocol is to formalize the parameters followed during the research. By following the procedure described in it, similar results should be obtained, considering a minimum variance due to the results of the algorithms the research platforms are based on and to possible future works, published after this research has been conducted. In the following

paragraphs it is explained how the three steps of the review process have been developed. The actual number of papers identified, accepted and rejected and the number of tools is reported in *Figure 1*.

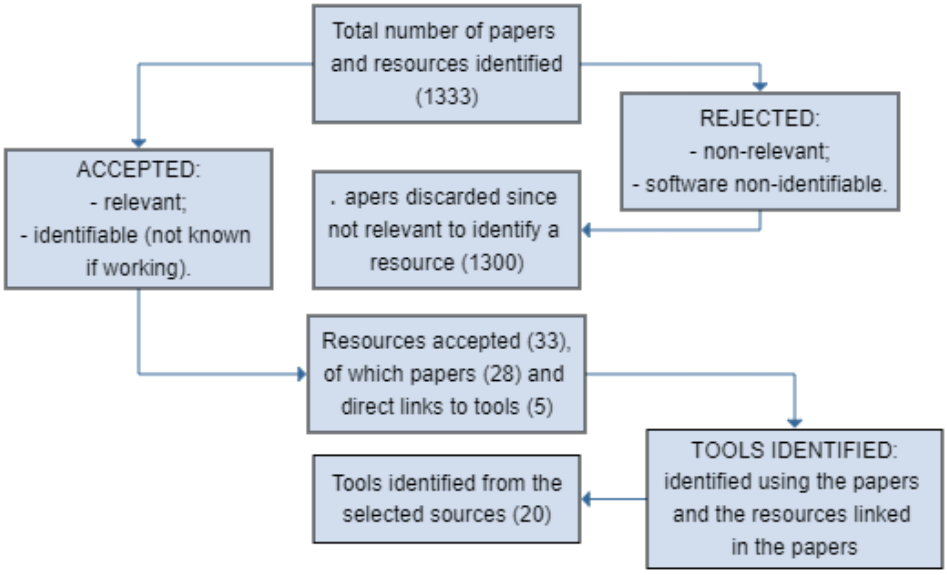

Figure 1. Review flow.

For what concerns the review there are some **pre-processing steps** to define before explaining it: searching strategy, starting set and acceptability criteria. First of all, a specific search approach has been adopted, the so-called snowball literature review method (Wohlin 2014). The snowball method, is a search strategy, based on the citations analysis, consisting in continuous backward and forward searches between one paper and the ones related to it through the citation network. Indeed, considering a starting paper, the forward snowballing consists in the retrieval of those papers where the current one is cited. The backward snowballing, vice versa, consists in the collection of the works referenced by the current paper, on the basis of its bibliographic references. The snowballing procedure is based on the union of these two phases, in a cascade mode. This same process is repeated on all the retrieved papers of a previous backward-forward step, in a flow in which the forward search, in a first moment, is carried out on the paper citing the current one, then on the paper citing the paper citing the current one, and so on. The same happens with the backward search. In order to avoid confusion about the order in the sequences of actions in the snowballing, a "waiting list" has been created. The papers accepted are inserted in the list and gradually digested through the review process once the ones retrieved before it have been processed too.

Second, in order to carry out a systematic literature review, there is the necessity to define some elements from which starting the process. These elements represent the starting point which the

citation network will be built on, through the snowball process. In this research case, the starting resources are two: a set of "seed papers" and a set of keywords considered relevant with respect to the research. The seed articles are resources identified before the beginning of the research itself. These are selected since they already are in the knowledge domain of the researcher (Lecy and Beatty 2012). Thus, the first part of the research is carried out by using this starting set. Indeed, the seed papers have a high probability of presenting, at least, a few relevant papers in their citation network. For instance, they may cite other papers as related works or, especially in case these resources are out since some time, they may be cited by newer studies as related works. Being a related work guarantees that the relevance with respect to the main work is high. For what concerns the keywords, these consist in a set of single or coupled words that are significant with respect to the research. At the beginning, it is provided a starting set of keywords through which navigate the literature network. Nonetheless this set will not necessarily remain unchanged during the research process, and, instead, some modifications can be made on it. Indeed, before the beginning of the review, the researcher may have some ideas about which may be key terms to get relevant information. Nonetheless, through the following analysis of actually relevant papers, new or different keywords may be identified and added to the original list. Alternatively, already present keywords may be refined with the aim of obtaining more focused results (Xiao and Watson 2019). Thus, the original set of keywords has been provided in the protocol (Cioffi 2022c), but at the end of the review it included two extra terms with respect to the starting set. Differently from the papers, the keywords provide indirectly papers which should be analysed before being accepted.

The third, and last, aspect of the pre-processing phase, indeed, regards the acceptability parameters for the papers (screening for inclusion). In order to be accepted the papers must fall within, at least, three parameters, stated before the beginning of the review process. These parameters are at three levels: generic, related to the screening procedure and to the full-text analysis. Each of them is applied in different moments of the review process. The generic one is applied at the time of selection, i.e. when a paper is identified in the citation network of another resource. It includes, as criteria, the language (English and Italian) and the publication date (from 2005 on). A paper can be accepted only if it is written in one of the selected languages, the two languages that can be reasonably understood by the researcher, and only if it has been published after the selected date. The aim of setting a temporal limit is to define a starting point before which it is highly unlikely that tools focused on this research topic have been created or that, if created, the tools are still maintained and working. These are the most generic parameters applied in this research and are used to quickly decide whether to keep or not a paper in a first phase. Instead, the part related to the screening phase considers a paper as acceptable only in case the paper's title, abstract or keywords include pertinent key terms. This

second step is applied at the time when the paper, after being accepted in the selection phase, is taken into consideration for being processed. In case one of the mentioned header parts contain at least one of the elements required, then the snowball is applied on the selected paper. Finally, the last parameter is the identification of relevant terms, expressions or tools in the paper full text. In this phase it becomes clear whether a paper is actually related to the main topic or not. This is the last parameter, applied in the second phase of the systematic review process, when the full text is finally taken into consideration. On the basis of these three parameters it is decided whether a paper can be considered or not in this research.

As concerns the **systematic review**, as previously anticipated, the process has been organized in three main steps, (see *Figure 2*). First of all, there is the actual collection of papers. Once it has been selected a paper to process, either as seed paper or retrieved with the keyword search, the process is the same for both the cases. In a first moment the paper is analysed in order to verify whether it should be accepted or not, then the citing and the cited sources are retrieved and analysed too. The keywords search is carried out on some selected openly available platforms: ACM digital library, EBSCO, Google Scholar, IEEE Xplore, Lens, ProQuest and OpenCitations. In particular for the case of Google Scholar, a fixed number of acceptable results, out of all the results returned by the research platform, has been selected. Indeed, in some cases the number of results for research is too high, in the hundreds, in order to be acceptable as such. Thus, it has been set a maximum of two hundred items as ceiling for the results to be reasonably accepted, on the basis of a study by (Xiao and Watson 2019) and a consideration by (Soulo 2019), for which this can be considered as a good point where almost all the relevant resources have been retrieved and from now on the results are statistically more non-relevant than the contrary. Nonetheless, in both these phases, the seed papers and the keywords ones, the backward and forward snowballing are carried out in the same way. Indeed, for what regards the backward search, i.e., the papers cited in the research, they were simply added one by one in a waiting list. This list was created in order to keep track of the next papers to analyse. Indeed, time by time each paper in this list is analysed in order to verify whether it may be useful to identify a tool or not. In case of affirmative answer, then the paper is accepted, and the snowball method is applied to it too. Otherwise, it is simply rejected and added to the list of rejected papers (created to keep track of the rejected papers too). Instead, regarding both backward and forward search, the DOI of the current paper, where present, is searched in the COCI REST API[2]. COCI, short for *OpenCitations Index of Crossref open DOI-to-DOI references*, is one of the possible ways to query the OpenCitations database, which allows to search for citing and cited sources. Indeed, by querying the COCI API it is

[2] https://opencitations.net/index/coci/api/v1

possible to obtain the works citing the searched source with the request ‘https://w3id.org/oc/index/coci/api/v1/citations/’ followed by the searched source DOI and, at the same time, it is possible to retrieve the works cited by the currently searched source with the request ‘https://w3id.org/oc/index/coci/api/v1/references/’ followed by the source DOI. The results obtained by querying the REST API for COCI show the papers citing the current one. In case these results are correct and not already present in the list of papers to analyse, they are added there. Instead, for what concerns the forward research only, in case no DOI is available, or in order to make research on different platforms than OpenCitations, the titles of the sources under investigation are searched on Lens, Google Scholar and ISI citation index. Again, in case the papers obtained through this research are accepted, they are added to the waiting list of papers to analyse. Thus, in a cascade mode, all the seed papers are analysed and together with them the citing and cited papers, and so on until no new paper can be retrieved with this method. This is the stopping rule of the process: once all of the papers found are analyzed, then, it is the moment to skip to the keywords search. The same cascade model is followed for the keywords since they are quite a lot and are searched on different platforms which provide different results. Again, this process can be considered as completed (i.e., stopping rule) in case no new paper can be identified following this method. At the end of this step, we skip to the application of the quality criteria in order to accept or reject the tools described by the up to now accepted papers.

The second step followed in the review process, is related to the quality assessment of the selected references extraction tools. This step and the following one, i.e. data extraction, are actually related and presented as contemporaneous in the protocol defined before the beginning of the research (Cioffi 2022c). In this step the papers and the related tools, are undergone to a set of criteria in order to define whether they should be kept with the perspective of being accepted or if, under a more analytic light, they should be rejected because not actually pertinent with respect to the research topic. Because of its structural relevance, this step can be defined either at the end of the previous step or iteratively before analyzing each single paper. In this way the process can be more complex but contemporaneously faster. This step is focused on the analysis of the tools and frameworks presented in the papers. It is aimed to verify whether they have been implemented as tools that can be used by a potential non expert user or if, instead, they are frameworks or toolchains not unified in a unique software. It is also investigated their availability on the web. The scope of this step is to filter out those solutions to the problem of references extraction and parsing, that are not based on actually available tools since:

- They are frameworks, i.e. sequences of independent steps not unified in a unique tool. This kind of solutions in most of the cases is the report on an analysis of possible solutions to the references extraction task, in a step preceding the creation of a tool. In some cases, instead, the framework is thought to allow the user to make adjustments to the single subparts of the framework, on the basis of the requirements of each specific task.
- In alternative, the process described in the papers is reported to be a tool, but this is not identifiable with an actually existing source. This is the case of tools for which no concrete reference is provided, e.g. the link to the resource or to a repository containing the software. Thus, since they cannot be found they can't be considered.

Instead, if the steps described in the paper are part of a unified tool which has an actual matching in the literature, then they are accepted for this phase, and in the last step their actual validity is checked.

Last, the data extraction is carried out at the end of the two previous steps. At this point the set of papers selected is definitive and thus, the related tools can be extracted and identified. But this step also includes a step over the identification. Indeed, the tools are analysed at a level deep enough to verify whether they can be actually used or not. This part of the research is based on the fact that all the possible tools, should be "valid" in order to be accepted. The adjective "valid" in this context can be associated to that of usable. The validity of the software and its consequent acceptability, indeed, are attributed on the basis of the before mentioned specific requirements, whose focus is on the actual usage possibilities of the tool. These criteria are: it is *available*, it is *compatible* with at least one of the best known operating systems (Windows, Linux and macOS) and it must be *openly available*, i.e. it must not be behind paywall. In the context of this third phase, the concept of *availability* differs from the one of *identifiability* presented in phase two. Indeed, in this case the tools are identifiable with an existing source, which is linked in the literature but it is not reachable, e.g. the URL does not point to the correct source, or working, e.g. it is not maintained. Vice versa, if the tool does not fit all of these parameters it is considered as invalid to the scope of the research and rejected. Once also this step has been completed the part of the review may be considered complete.

The tools resulted from the review, both the valid and invalid ones, are presented in *Table 1* with a brief description of the tools reported. For what concerns the online tools the validity can be a parameter not valid for future research. Indeed, while at the beginning of this analysis the tool PDFdigest (Ferrés et al. 2018) did not work, after three months it started working again. So, for a matter of correctness they are reported together with the valid ones, with a specification regarding the validity status, in case in the future they will be valid again. The total number of identified tools after

the second step is of twenty, but at the end of this third step they are reduced to twelve because of the invalidity of eight tools.

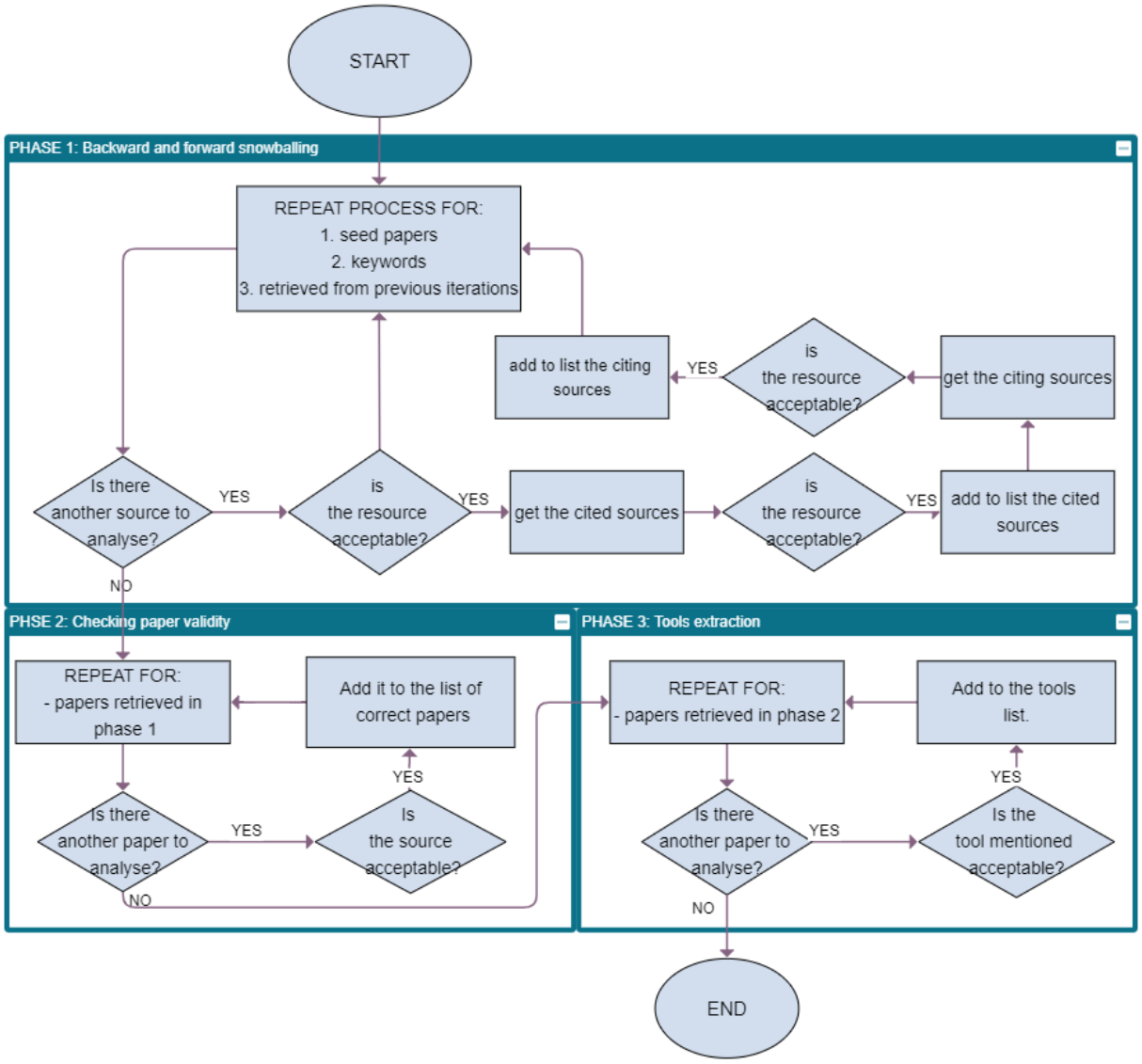

Figure 2. Sequences of action of the literature Review

2.2 Relevant Tools for References Extraction

As anticipated, the software selected at the end of the literature review are twenty. All these tools are based on different text extraction methods. Indeed, many different techniques have been developed over the years in order to carry out, in general, the text, and, more specifically, the references extraction task. These methods have given birth to plenty of tools which either use one single method or, more often, a combination of methods. The most used methods are machine learning, rule-based

methods (including regular expressions based methods) and hybrid methods (Hashmi, Qayyum, and Muhammad Tanvir 2020).

One of the most widespread methods is *machine learning*. Machine learning based methods can be either supervised or non-supervised, but the supervised version is normally preferred. In this kind of methods, the procedure is based on training a model that will be later used in the actual extraction and parsing tasks. Both these tasks are carried out following the strategies for the sequence tagging problems. Indeed, the input is considered to be composed by objects with features. These objects, which can be identified with the text sections, figures etc., are extracted following this process: first, the text is split into small sequences (tokens), the tokens are labelled and, finally, the tokens with the same labels are grouped. The grouped tokens represent the text objects that now can be extracted. Among the supervised machine-learning algorithms, the Conditional Random Fields (CRF) is the most used one. Anystyle[3] is based on a machine learning structure, in particular, but , CRF. Even if in its online version it is only available to parse single references, it is also available a gem which provides also bibliographic references extraction and parsing from full texts PDF files. CEBBIP (Gao, Tang, and Lin 2009*)*, a parser for Chinese electronic books, is one of the tools which makes use of CRF in combination with rule-based system and clustering. Each of these methods has a specific scope: the rule-based system to locate and parse references in the PDF, the CRF to extract and the clustering to enhance the parsing phase. CERMINE (Tkaczyk et al. 2014; 2015) is also a CRF based system for parsing the bibliographic references, but which also makes use of Support Vector Machine (SVM), another machine learning algorithm, for text zones classification and K-means clustering algorithms for reference string extraction. The ExCite toolchain includes the use of Cermine in conjunction with ExParser (Hosseini et al. 2019). The latter is a tool implemented by the ExCite project, aimed at references extraction and parsing, and based, again, on CRF. GROBID, differently, is a tool almost only based on CRF systems in order to carry out all of its extraction tasks (Romary and Lopez 2015). Indeed, among the challenges carried out by GROBID, there are also the header metadata extraction and the labelled text sections extraction. Parscit (Councill, Giles, and Kan 2008) is a tool aimed at extracting and parsing references. In a preprocessing step it converts the files from PDF to text, it then extracts the references through the use of a set of heuristics and parses them one by one with the CRF system. Finally, it returns them in the selected output format. RefExt is again a tool provided by the ExCite project (Körner et al. 2017). Similarly, to the main implemented ExCite toolchain, its functioning is based on Cermine. Indeed, Cermine provides the images of the papers that are used as input for RefExt which, with the layout information, get the position of the references,

[3] https://github.com/inukshuk/anystyle

extracts and parses them. Finally, Science Parse[4] is the last of the tools reported here which is based on Conditional Random Fields. It first sends the single pages of the input PDF to PDFBox[5] and then, on the resulting files it is applied the CRF based methodology to extract the tokenized references.

Differently, other tools implement *rule-based methods*. Using sets of rules to extract sections of text is based on the concept that all of them have different features. The rules are a formalization of these textual (e.g., font style, font size) and layout features that are used to identify the sections and possible subsections. BRExSys, and its specific application, DeepBiRD (Rizvi, Dengel, and Ahmed 2020), make use of a deep neural-network method to recognize the references section. Indeed, in a first step, the PDF is seen as a defined image, which is then feed to the neural network which recognize the references by the identification of the single image pixels. Differently, CITEREP implements a rule-based system to extract a particular metadata from the bibliographic references, the journals (Verkuil 2016). In order to do this, it first identifies the bibliographic section, then it extracts the references and, finally parses them in order to retrieve the journal. PDFdigest (Ferrés et al. 2018) is a tool aimed at extracting the textual and layout features from PDF files. After having extracted the information, it returns them in XML format. To extract the references, it makes use of layout information. Pdfextract[6], similarly, works on the papers' layout. By identifying the references regions and the single references inside it, Pdfextract can extract and parse the references with one of the tools provided by Crossref.[7] Finally, PDFX implements a union of rule-based methods aimed at extracting the input PDF structure and baseline (Constantin, Pettifer, and Voronkov 2013). Through the identification of the structure, it can extract the sections together with their features, including the references one. Once identified the reference, with the same procedure, it also extracts the single references and their metadata.

A specific typology of rule-based methods implemented to extract the bibliographic references is based on the *regular expressions* (or *regex*). The regular expressions are sequences of symbols that are used to match string patterns. In this case, thus, the regex are used to match the typologies of strings identifiable with the references. This is a simpler method which usually comes with a pre-processing step. In the case of PDFSSA4MET[8], the PDF file is previously converted in XML by pdf2xml[9] and lately the references are identified. OCR++[10] is a CRF based tool, which instead, only

[4] https://github.com/allenai/science-parse
[5] https://pdfbox.apache.org/
[6] https://www.crossref.org/labs/pdfextract
[7] Simple Text Query system or Crossref Metadata Search, reachable at the web page http://search.crossref.org/
[8] https://github.com/eliask/pdfssa4met
[9] https://sourceforge.net/projects/pdf2xml/
[10] http://www.cnergres.iitkgp.ac.in/OCR++/home/

for the references extraction task makes use of regular expressions. In the related paper (Singh et al. 2016) the full set of 16 typologies identified is provided and explained. In this way, while it identifies the sections with the CRF, it separately extracts and identifies the references only through the specified regex. The parsed references are finally returned in XML.

Another typology of tools is the hybrid typology, which can be defined as a "tools suite" or toolchain. In this case there is not one single tool carrying out the full task but, instead, there is a tool which, through the combination of different tools, one for each sub steps of the main task, is able to do it. An example of this typology is PDFMEF (Ning, Jin, and Wu 2006) which allows for the extraction of almost all the information available in a PDF file, including images and mathematical formulas using specific tools, i.e., PDFigures2 (Clark and Divvala 2016) and a Java jar file respectively. Also, RefUTU (Holvitie and Leppänen 2015) can be defined as a tools suite since it combines in a unique tool two different existing tools. These two tools are used following the processes of conversion to plain text and extraction of the bibliographic references, PDFExtract and Freecite.[11]

Apart from the previously mentioned methods, there is another typology of tools based on the use of an API to extract the references. The only tool which makes use of it is Scholarcy[12]. It is a recent tool, which through the use of the API is able to parse and match the references' parts with information available in other databases.

All these mentioned tools are reported in the table below together with their relevant characteristics. These are: the validity status with the explanation for the potential invalidity, to identify the ones which could potentially be used in order to carry out this research and which not, the creator or the creating group of each tool, the features that the tool is capable to extract (the references and possible other parts of the texts) and the system on which the tool is based.

[11] https://github.com/miriam/free_cite
[12] https://www.scholarcy.com/about-us/

Table 1. List of all the tools. If their status is 'invalid' one of these labes is attributed: "NI" no resource can be identified, "NA", it is provided a link to the resource but a server error is returned; "NM", the tool exists but it is not maintained.

Tool	Creator	Approach	Extracted Features	Validity (at the moment of the review)
Anystyle	Keil	Machine learning (CRF etc.)	References	VALID
BRExSys (DeepBiRD)	Rizvi et al.	Rule and image based	References	INVALID (NI)
CEBBIP	Gao et al.	Rule-based, machine-learning (CRF), clustering	References	INVALID (NI)
CERMINE	Tkaczyk et al.	Machine learning (CRF, k-means, SVM)	Metadata, references, text sections	VALID
CITEREP	Verkuil	Rule-based	References	VALID
Dr. Inventor	Ronzano et al.	Text mining tools and on-line services	References	VALID
ExCite	Boukhers et al.	Machine learning (CRF etc.)	References	VALID
GROBID	Lopez et al.	Machine learning (CRF)	Header, sections, references	VALID
OCR++	Mayank et al.	Regular expressions	References	INVALID (NA)
ParsCit	Councill et al.	Machine learning (CRF etc.)	Author affiliation, section labeling, references	INVALID (NM)
IceCite	Bast et al.	Rules and regular expressions	Header and references	INVALID (NA)
PDFdigest	Ferres et al.	Rule-based (layout analysis)	Sections, references	VALID
PDFExtract	Crossref	Rule-based (layout analysis)	References	INVALID (NM)
PDFMEF	SeerLab	Hybrid	Header, sections, tables, mathematical expressions, figures, references	VALID
PDFSSA4MET	Kunnas	Regular expressions	References	VALID
PDFX	Constantin et al.	Rule-based	Title, tables, sections, references	INVALID (NA)
REFext	ExCite project	Machine learning (CRF etc.)	References	VALID
RefUTU	Holvitie et al.	Hybrid	Header and references	INVALID (NA)
Scholarcy	Gooch et al.	API	References	VALID
Science Parse	AllenAI	Machine learning (CRF etc.)	Sections, references	VALID

The ones presented above are all the tools that after the literature review are valid in order to be compared on the basis of the references extraction task. The validity status is a checkpoint, and the tools which do not pass this parameter cannot be considered in the research. Nonetheless not all the tools selected up to this moment and considered valid will be used in the review. Indeed, further requirements for the tools will be specified in the methodology section and that will be the third and last step through which the tools will be passed to be used. The reason for putting further parameters is that, to understand at a concrete level if the tools can be used or not, they must be downloaded and tested before understanding if they do actually work. To sum up the steps carried out on the software in order to select them during this systematic review and the one that will be carried out in the methodology, there have been three steps in which the tools have been selected:

1. General criteria (during the literature review): relatedness of the tools' features to the main research topic.
2. Validity status (during the literature review): the software can be retrieved somewhere and is apparently usable.
3. Further parameters required for the software to be accepted, regarding its concrete usage (explained in the chapter "Methodology").

2.3 Approaches and Frameworks to References Extraction

Apart from the tools identified there are some tools and workflows which can be interesting to consider before the beginning of the research. Indeed, even if they are not considered as part of the research because they are not exactly centered with respect to the research topic or they are sequences of actions not unified in a single tool, some topics and procedures are interesting with respect to the research topic. The related tools can be classified on the basis of their specific features: tools for parsing single references, tools for parsing references lists, workflows for extracting and parsing bibliographic references form the PDFs and metadata extractors from PDFs.

Single references parsing. This category of tools is relevant in the context of citation parsing. Indeed, it represents a set of tools which is able to parse single references and return the metadata they are composed of in structured form, either BibTeX, XML or other formats. Inside this category the tools can be really different on the basis of the building system, the input data they accept, the focus on different typologies of citation, or on the ability to extract a different number of metadata form the reference strings. Some of these tools are based on machine learning techniques. It is the case, for

instance, of Freecite,[13] a CRF-based web app for parsing citations; also, Reference tagger,[14] again a tool based on CRF for parsing academic citations. Moreover, Xiaoli Zhang (Zhang et al. 2011) introduces a tool which makes use of a different machine learning technique, based on SVM in order to parse references. Differently, (Hetzner 2008), (Yin et al. 2004) and (Ojokoh, Zhang, and Tang 2011) illustrate three different methods based on Hidden Markov Model to carry out the same task, i.e. simple HMM, bigram HMM and trigram HMM respectively. Other tools are based on different techniques. Some instances of remaining building typologies are: Biblio,[15] a Perl library which makes use of regular expressions, BibPro,[16] a tool based on sequence alignment and the work presented by (Suryawati and Widyantoro 2017) in which rule-based, heuristics and machine learning are combined in a unique strategy. Similarly, to BibPro, (Hsieh et al. 2014) propose frame-based approach to extract references. In this work it is reported this tool to outperform CRF based methods for this task.

Parsers for references lists. This is a category of tools able to extract and parse references from files in different formats, but not from full texts pdf files. Indeed, in most of the cases they can, given a text file with a list of references, possibly where each line corresponds to a reference, extract the single references, parse them and return the metadata of each sentence. This category is close to the previous one. Indeed, both these tool typologies are based on the concept of parsing the references out of their original context, the full-text PDF paper. The main difference is the fact that the previously described tools are able only to parse single references and not references in text blocks while the current ones can take as input both single references and blocks of references. It is the case of Citation,[17] a regular expression- and rules-based tool for parsing citations in citation lists; Citation-parser,[18] a rule-based parser; Neural Parscit,[19] a deep-learning tool based on LSTM, Long Short-Term Memory, created as branch of ParsCit; and Refparse,[20] a tool written in Java, based on regular expressions, for parsing citations in list in XML format.

Frameworks for parsing bibliographic references in PDF full text. A different perspective on the bibliographic references identification, extraction and parsing tasks is provided by many workflows proposed during the years. These solutions are presented as related works since they are composed by a sequence of independent and separated actions, not unified in a single tool like the research topic

[13] https://github.com/miriam/free_cite
[14] https://github.com/rmcgibbo/reftagger
[15] http://search.cpan.org/~mjewell/Biblio-Citation-Parser-1.10/
[16] https://github.com/ice91/BibPro
[17] https://github.com/nishimuuu/citation
[18] https://github.com/manishbisht/Citation-Parser
[19] https://github.com/WING-NUS/Neural-ParsCit
[20] https://github.com/VBRANT/refparse

requires. (Peng and McCallum 2004) describe a machine-learning based framework, which outperforms the results obtained on the same input dataset by an HMM based method. Similarly, (Tkaczyk et al. 2014) explore a composed tool based on simple HMM and rules thought to be easy to be modified by the user. Other solutions of this type are based on rules. (Azimjonov and Alikhanov 2018) introduce a methodology for data and bibliographic references extraction based on fixed rules. (Huynh and Hoang 2010) introduce the GATE research framework, by combining the layout information to a rule-based system. Also (Kluegl, Hotho, and Puppe 2010) funds his method on a set of rules. The idea is to allow the framework to adapt to the different journals or papers contexts to have a higher performance. Other solutions to this topic are represented by the frameworks Semrex (Ning, Jin, and Wu 2006), based on the use of ontologies, and DeepBIBX (Bhardwaj et al. 2017) a framework for parsing references from images using deep fully convolutional networks.

Other typologies of pdf extraction tools. One different typology of pdf extraction tools is represented by those tools capable of extracting the reference metadata from the header of the papers they receive in input. CB2BIB[21] and Mendeley[22]. Nonetheless, they are not able to extract the bibliographic references of the papers. One last unicum in the panorama of the extraction of cited resources is represented by PdfX.[23] This last tool, based on regular expressions, is able to extract the links to cited web resources and return them as output. It also provides the possibility of automatically downloading the linked resources.

[21] https://www.molspaces.com/cb2bib/
[22] https://www.mendeley.com/
[23] https://github.com/metachris/pdfx

3. Methodology

In this section it is explained the methodology followed from the selection of the definitive references extraction tools to the quality evaluation of the references extracted by each of the tools. The methodology is described separately from the related data, that will be presented in the next chapter. The steps followed in this methodology originate from the identification of the working tools at the end of the literature review. These steps are:

- The first step consists in the definition of a set of rules aimed at selecting the references extraction tools relevant for the research. Any software which does not accomplish those requirements are rejected and considered as related works.
- Once the final tools are identified, the next step is based on the selection of the set of papers to use as input to test the references extraction tools. Then, starting from the original dataset, it is created a gold standard, or ground truth, related to it. The gold standard is a formal representation of the input references against which to test the tools results in order to verify their quality. This step's objective is to provide the materials for the final comparison between the output of the parsers and the reasonably best results which could be extracted by the dataset papers, i.e. the gold standard.
- The third phase consists in converting the output files of the extraction tools to the language selected for the ground truth results. This step is thought in the perspective of the final comparison. In this way it will be easier to compare the values of the two results if they are written in the same language.
- The final step of the methodology consists in the identification of the parameters for assessing the similarity between the ground truth and the output files and implementing them. After the last step the procedure is carried out, the comparison takes place, and the results are analysed based on the dataset and parsers' peculiarities.

These steps can be generically divided into two categories. The first two steps, i.e., definition of the software requirements and dataset creation, lay the groundwork for this study. The last two steps, instead, i.e., conversion to gold standard language and evaluation, take as input the data obtained in the two previous steps and, by working on and combining them, obtain the final results. These steps are reported in *Figure 3*, together with their main steps (continuous thick line) and pre-processing steps (stroke thin lines).

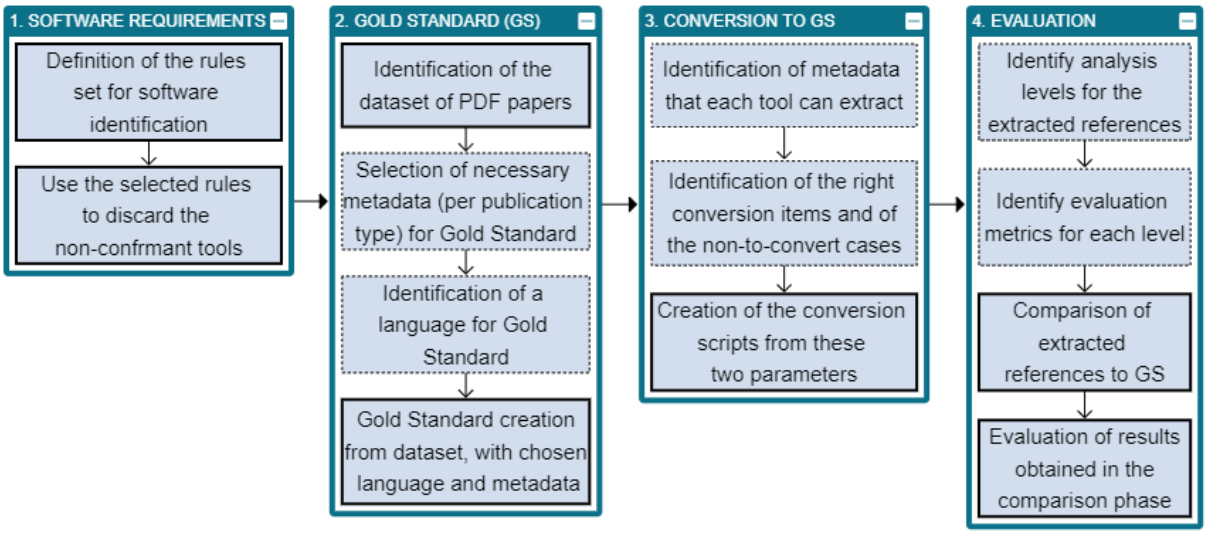

Figure 3. Methodology flow

3.1 Software Requirements

As shown in the previous chapter, *2. Literature Review*, the references extraction tools can be classified on the basis of their approach to the extraction task. At the same time, the aim of this research was not to select the tools on the basis of their approach to the references' extraction task. Instead, all the approaches are accepted if the tools undergo some defined parameters. These parameters have been selected keeping into consideration the needs of the target user considered for this research, i.e. the small publishers. The following list, apart from newly created parameters, also includes some of the parameters used in the last step of the literature review. Indeed, its role in this research wants to be a full and self-conclusive set of parameters, following which it is possible to retrieve all the usable tools starting from the literature review. These parameters can be applied to the tools that have been retrieved from the second step of the review on, i.e. it is a unique tool that can actually be retrieved somewhere on the web. In this way, it is possible to have a complete overview of the tools' selection process. For the purposes of this research only the extraction and parsing tools which responded to the following features were taken into consideration:

1. The *focus* of the tools must be on the *referenced sources*. This first step can be more specifically considered as a preprocessing step. Indeed, in some cases from only reading the article and making preliminary tests it is not clear whether the tool is able to parse the references or only the papers metadata, e.g., title, authors, venue. Thus, before starting the

actual exclusion of the non-appropriate tools it is necessary to verify whether the tools are compliant with the major task of the research.

2. It must *parse full text PDF papers*. The aim of the research is to identify those tools which can take in input a full text research paper or a generic PDF file and extract its references. This can happen either by identifying the bibliographic section or by looking for the references list as such. Nonetheless, it is essential that the extraction tools can produce results starting from texts, in PDF format, not pre-processed with other tools with the aim of parsing the bibliographic references. It is not necessary that the software can parse scanned PDF documents: the occurrence of those papers in recent years is scarce, since normally the papers are written in digital form since the beginning.

3. It must retrieve *singularly tagged references*. The first and less specific degree of precision in the reference extraction task is the necessity to extract the references as single entities and not as a group. Indeed, some tools can extract the references as a block inside the full text paper. But to get the references with precision it is necessary for the tool to extract the references as single entities. Thus, the presence of one or more ways to recognise the point in which one reference ends and the next one starts is mandatory.

4. It must retrieve the *metadata of each reference*. At a quite specific level of precision, it is necessary for a tool to be selected that the metadata of each reference identified are recognised. Indeed, it is not enough that the single reference is identified as whole. Its inner metadata should be recognised, otherwise the content of the reference is still an only human readable piece of text. For a reference to be structured, enough metadata to qualify it must be identified. For what concerns this last point, it is not necessary that all the present metadata are identified by the extraction tool. Nonetheless, enough metadata for each reference should be reported. The range is that of metadata which are reasonably present in all the references, i.e., title, year, and in a quantity enough to identify each reference in the bibliography.

5. It must be either an *application or a ready to use programming language* (e.g., Python or Java) *library* (e.g., no training required by the final user), better if usable by the command line. This requirement is based on the target user rather than on the search per se. Indeed, the user is expected not to be expert in some specific programming language, thus an application could be the best option since it allows for the most simplicity of use. Nonetheless, also some libraries have been created in different languages which require only a minimum effort to be working. Also, this kind of structure is acceptable for a matter of usage. Different tools, less

structured and less usable by non-expert users, are not considered since they would require too deep informatic knowledge which is not expected by the target user.

6. It must be *working*, in case of libraries or locally downloadable applications, or *reachable*, in case of only online usable applications (i.e., the website must be up). This requirement is voted to guarantee that the tool is working and that it is freely available to use. The tools must be, following this requirement, existent, thus a name or a repository link must be provided, working, thus it is well maintained, and open source.

7. It must be an *independent software*. This last requirement is stated to avoid the presence of suites which simply reuse other software without providing any specific improvement.

From the literature review, it was retrieved a wide specter of extraction and parsing tools which filled at least one of the previous requirements. Nonetheless, only a few of them matched all the requirements and then, were effectively considered in the research. Indeed, the relevant tools identified for this research are 7, starting from a base of 20. All the other tools have been rejected because of at least one of the previous parameters. In the following chapters the tools excluded by one of the above-mentioned categories have been used as example in order to explain the role of those categories.

The first requirement allows to reject a small number of well-known tools. This typology is majorly represented by tools focused on the categorisation of the scientific articles in user-created libraries. Their main objective is to retrieve, format and make available the metadata of each of the added papers and then being able to retrieve them in case they are needed. The two tools which can best represent this category are CB2BIB[24] and Mendeley Desktop. The aim of these tools indeed, is that of retrieving the metadata of the tools, and, where appropriate, their section, but they cannot extract the works cited in the text or in the bibliographic section. Thus, even if those tools are well known and working, they won't be considered in this research.

The second typology of works rejected for the scope of the current research regards those tools which are not able to parse full-text PDF files. This kind of tool is only focused on the citations parsing task but is not able to work directly with the PDFs. In this set of tools are included for instance BibPro (Chen et al. 2012) and SciWing (Ramesh Kashyap and Kan 2020). Those two are representative of two different kinds of tools. On the one hand BibPro is a tool focused on the reference parsing activity per se. It is not able to take a full text and return the parsed strings since its scope is that of, by taking

[24] https://www.molspaces.com/cb2bib/doc/overview/

as input a reference string and parse it, with a sequence alignment technique. Another tool like this is Citation[25], which only accepts single strings, ready to be parsed, as input, and Neural ParsCit, which instead accepts both single references and text files containing sequences of references that are parsed singularly. On the other hand, SciWing is a more complex tool, derived from ParsCit. Its scope is wider than the previously mentioned tools and it has the ability of taking a full text and parsing it. Despite this positive aspect, SciWing does not take PDFs as input files[26]. Instead, they are still based on text files. Thus, a PDF file should be pre-processed with other tools, e.g., PDFBox, before being parsed by the main tool. Since this is out of the scope of this research this kind of tools are necessarily excluded.

The exclusion of a third software's category concerns the fact that they are not compliant with the concept of tagging the references as such. Indeed, while almost all these tools take as input for being processed full-texts files in PDF format, they have been rejected because they are not able to parse the single references. Inside this category, the most frequent tool typology is the one having as objective the identification of the paper sections. While this kind of analysis is deeper than the one presented in the previous point, it is still not enough to be acceptable for the purposes of this research. An example of this selection is the Parscit branch SectLabel (Luong, Nguyen, and Kan 2012). The tools, built like this one, are aimed at identifying the papers sections through the analysis of the paper layout. Indeed, the text is considered as a sequence of lines which are analysed through CRF and OCR systems. Then, the sections are identified based on their title. Nonetheless, since the SectLabel level of analysis stops at this point, it cannot be considered enough to be considered as a reference extractor.

For what concerns the fourth parameter in the required parameters list, two different categories are recognisable: tools which are not able to identify the metadata at all and tools which are not able to identify enough metadata for each reference. Indeed, the former case is represented by a set of parsers able to retrieve the single references but not their inner metadata. It is the case, for instance, of OCR++ (Singh 2016) and of PDFdigest (Ferrés et al. 2018). Both these tools are aimed at identifying the sections of the papers, and for both of them identifying the single references is the deepest level of analysis for the "References" section. Thus, while the references section and the single references, which the section is composed of, are identified, the inner metadata, composing each single reference, are ignored. The latter set instead, is composed by tools which can recognise some of the references metadata, but not enough to define them or not available in all the references. It is the case of PdfX

[25] https://github.com/nishimuuu/citation
[26] https://sciwing.readthedocs.io/en/latest/usage/tutorials.html

by Chris Hager[27]. Indeed, this tool is able to retrieve only the URLs. While this metadata can be enough to recognise a referenced source, it is not known in how many references this metadata is present. Surely, not in all of them. Thus, if only one parameter is considered for the identification of a reference, it is not an affordable tool for the identification of all the references. The same thing can be said for CiteRep. Indeed, this tool has been thought with the perspective of extracting only the journals from the references section. With the aim of this project, the journals metadata only are not enough to let the paper being accepted in this review.

For what concerns the usage of the software, it was considered as relevant the availability of the software, i.e. it exists somewhere. Also, in this case there are two sub-cases falling into it. The first one regards extraction tools not identifiable with a concretely existing resource. There are indeed plenty of works describing processes or tools, which cannot be replicated or retrieved. Indeed, in the first case they usually represent only a sequence of passages, not unified in a single tool. And a second case regards those papers which are presented and described in one or more papers retrieved in the literature review, but they cannot be found anywhere, because no link to that resource is available. It is the case for instance of BRExSys and CEBBIP. Both are described by one or more papers but a link to a resource is not provided nor is it possible to find it on the web. Thus, since it was not possible to verify their life and usage status. Nonetheless, these tools should have been excluded since the literature review step. A second typology of software not accepted in this research because of its usage is the one which includes parsers which are expected to be trained by the final user to work. Also in this case, on the one hand with the perspective of a non-expert use, and on the other for the lack of training materials in this research because of the variety of fields tested, this typology of tools could not be considered. An example is the software Refext, provided by the ExCite project. Indeed, the tool has mandatorily to be trained before the first usage in the local environment. Even if the models are ready to be used in the original repository, however, the level of difficulty to work with it and the time required to prepare the tool is high.

Based on the sixth parameter, requiring that the tool is actually working, a particular category of software was excluded by the research. This last one, was not derived by a structural issue, but rather for a concrete one. Indeed, this section includes all the tools that were originally selected to be studied in the research but that in a second moment came out that they cannot be currently used. The reasons vary on the basis of the typology of software. Some of them are available only as an online service and the link to that service is not working. It is the case of RefUTU and PDFX. Other tools have gone out of usen and are no longer considered as active. Two examples of this issue are PDFExtract and

[27] https://github.com/metachris/pdfx

Semrex. In particular, the PDFExtract authors themselves, in order to help the users, provide the name of a substitutive tool which in their opinion was the best option as substitution: CERMINE. Finally, some extraction tools have not been totally gone out of use, but the same authors have produced new tools aimed at substituting the older one. The best tool which represents this occurrence is Parscit. This aspect comes out in different ways from the tool GitHub repository. On the one hand it is reported: *"**Update**: While we continue to partially support the codebase, we highly recommend you to use our neural version of the reference string parser here: https://github.com/WING-NUS/Neural-ParsCit*". But the confirmation of the complete dismissal of ParsCit comes from the fact that CRF++-0.51, one of the dependencies required by the developers, is a version which does not work anymore. In one of the issues, specifically asked about this fact, one of the creators suggests not to use Parscit in its non-neural version.[28]

Finally, the last category of excluded tools regards the software suites. With this term reference is made to the tools that coordinate other tools to carry out different tasks at the same time. It is the case of PDFMEF which, using different PDF parsers, carries out contemporaneously different extraction tasks. While this kind of tool can be useful to extract different features from the papers, with the most appropriate tools for each specific task, PDFMEF per se does not add any contribution to the extraction and parsing task. These kinds of tools indeed, while being useful in order to parse full texts, are simply coordinating other independently functioning tools. Thus, the inner parsers rather than the software suite should be evaluated.

The remaining tools selected during the literature review process have been considered for the purposes of the research. They are introduced and described in the next chapter,
4.1 References Extraction Tools.

3.2 Gold Standard

The next step in the task of comparing and evaluating the parsing tools regards the preparation of the data to use to test the parsers qualities. This task can be defined as the union of two subtasks. The first subtask consists in the identification of an initial dataset of papers in PDF format to test the references extraction tools. The second subtask instead concerns the actual creation of the gold standard from the input dataset.

[28] The current citation is taken from https://github.com/knmnyn/ParsCit/issues/35. knmnyn **commented on 29 Jul 2021: "**Hi ParsCit is not regularly supported anymore as we have more recent toolkits that use neural approaches that outperform this method. You're welcomed to check out NeuralParsCit or SciWING as replacements."

The former subtask, the selection of a set of PDF files, is the first step in this work that does not directly regard the parsers. Indeed, the scope of this task is creating a dataset of papers in PDF format that will be used as input for the tools whose aim is parsing them and providing their respective structured references as output. However, before identifying the input papers, there are some parameters to take into consideration. First of all, it is necessary to identify a set of PDF papers being representative of the two main specific characteristics required for this research: one publication type and different research fields. Indeed, on the one hand, for the purposes of this research the tools need to be tested with one specific type of publication, i.e., the research papers. On the other hand, the tools extraction and parsing qualities must be tested in a variety of research fields. In fact, the aim of this research is to test the quality level of each different software selected against a wide different selection of fields (see *Table 2.* Research fields, related short names, papers DOI and short papers names), in order to understand if some of them are more prone to some specific types of errors or if, instead, they are generically good or bad at carrying out the task, independently from the field. Indeed, each research field has different types of publication standards, references styles, and even publication types cited. For instance, papers belonging to the scientific field are more likely to cite journal articles than books, while in the social sciences the number of books and reports cited is higher. Moreover, these differences are not just a matter of the single fields, but even inside the same citation fields there are different journals providing different paper layout, citation styles and rules. For instance, two different journals may suggest using the APA style but one of them specifies the journal title and the other does not. Thus, the selection of the papers must be random within a wide set of different publications coming from different research fields and published by different editors. The number of papers selected for this research is not high with respect to the mean number of papers selected in similar types of research. Nonetheless, the number of papers, even if small, must be representative of the different aspects described above.

The second subtask is still related to the PDF input dataset, but it is more articulated. The gold standard is a baseline step in which it is created the ground truth that will be used as a comparison metre in order to evaluate the tools capabilities to extract the bibliographic data from the PDF papers together with the metadata they are composed of. The idea which lays behind the creation of a gold standard dataset is to evaluate the results against a given dataset, created separately. The term "gold standard" refers to a benchmark, representing the best version of a test under reasonable conditions. Which can also be said as: the gold standard is the reasonably best output that can be reached by means of modifying an original dataset following the principles of the research. In the current research the gold standard is necessary to quantify the number of references correctly extracted (or not) by the selected tools from the previously identified set of PDF papers. Since the objective of the research is

comparing some selected references extractors and parsers based on their results in the task of extracting and parsing the references contained in full-texts PDF files, the gold standard will be the best representation of the references that can be extracted from the input PDF files in the dataset. From a concrete point of view, the aim of this process is to provide a set of files, corresponding to the original dataset, which present only the references in a structured format. In order to create it, the bibliographic references must be manually extracted and written in the selected language. In this way it is guaranteed the highest possible level of correctness in the identification of the citational data since the references are provided in natural language in the input files. One final consideration about the creation of this ground truth is that the references should be written in a language that allows a comparison between the results of the extraction tool and the files themselves. Indeed, the final objective of the gold standard is to create a testing set of citation files against which to compare and evaluate the results of the bibliographic references parsing process.

In order to create the gold standard, two pre-processing steps were necessary. The first step consists in the identification of the metadata which are considered as necessary for the description and identification of the publication typologies. Second and last step, there is to identify a language and, possibly, a vocabulary capable of representing the citational data in a structured way. Only after the identification of these aspects, the process of the gold standard creation can take place.

The first preprocessing step carried out is the selection of the metadata considered in the gold standard. This passage is based on the results pointed out in the article (Santos, Peroni and Mucheroni 2022). To select the metadata to consider in the gold standard, the average metadata identified for each kind of publication type has been selected, as they are reported in the aforementioned article. The scope of this selection is to identify a tool which can retrieve, at least, the metadata which are standard for each specific type of publication. Therefore, it is not expected the perfect extraction of all the metadata available in the references, for the purpose of this research. In case the tools can extract more metadata than the ones considered, this aspect won't provide a higher score to the software. The aforementioned article has been followed literally for the organisation of the standard metadata. Nonetheless, there is an aspect which has been managed differently from the reported data. Indeed, in the original article, in cases in which a metadata of the same type (e.g. the journal title) is present in two research fields presenting formal differences (e.g. full or abridged format), the author does not consider them as standard metadata for the publication type. Instead, in some limited cases it seemed senseful to consider the metadata even if the shape is not the same. For instance, in the case of the journal articles the journal title is present in all the research fields considered but, while in some

cases it is reported in abridged format, in others it is as full format, but in both cases the metadata (the title) is present. Therefore, the metadata is considered by accepting both the formats.

The second preprocessing step concerns the selection of a language to use in order to create the gold standard references-only files. The language has a relevant aspect at this point. Indeed, it is the means through which the references metadata are carried. Thus, it should be selected a language capable of conveying all the semantics required by the selected metadata with a certain range of elasticity. For instance, there are languages in which for each metadata there is a specific type of tagging without the possibility of adding further information. Differently, others are more likely to be provided with addings or attributes or different levels of description which provide a higher level of accuracy and at the same time the possibility for less specific taggings in case of uncertainties. Once identified the language, the relevant metadata should be associated to the type of encoding selected. Then it must be applied to all the dataset papers references. At this point the gold standard has been created and kept apart for the comparison phase. The actual gold standard will be presented in the chapter describing the data used in the thesis, in the next chapter.

3.3 Conversion to Gold Standard Language

This step regards the conversion of the files obtained as output of the extraction and parsing tasks for each selected tool, from the language in which the tools has structured them to TEI XML, the language of the gold standard. This step is required to let the comparison take place since only by comparing two texts written in the same language it is possible to make a comparison among them. Indeed, in this way, it is possible to create one single file script for the comparison between the output (in TEI XML after the conversion) and the gold standard texts (TEI XML). Concretely, the conversion phase coincides with the creation of script files for the conversion of the format in which the extracted references have been coded to XML TEI. Indeed, all the tools for references extraction and parsing retrieved in this research provide the output citational data in different formats. Jats, BIBTEX and plain text are the most common file formats used to code the texts. Often more than one option is available, and the final user has the choice to select the preferred output format.

This step has some critical aspects. First of all, as with all natural languages, machine readable languages have differences in terms of expression between them. This aspect is derived from the different necessities for which the languages are created. Indeed, while XML TEI is able to describe all the written materials, since it has been thought to be used by libraries and archives, BibTeX was

thought only to express bibliographic materials and, therefore, the possibility of expression is limited to that field. Also, the development of the languages is another element to consider: the more and the wider a language is used, the higher it's possibility to express different concepts. This consideration was necessary in order to understand how to act during the conversion. For instance, Cermine, even if using Jats, makes use only of a small subset of its XML tags (see *Table 5*). Therefore, when converting the files from Jats to XML it was necessary to understand which elements were not tagged, which others were badly tagged and how to consider and translate these tags. The specific elements will be discussed in the specific section dedicated to the single tools' conversion scripts.

A second and last issue to consider, before creating the conversion script files, regards the selection of metadata that should be converted and the ones that should not. Indeed, by analysing some of the data obtained as results after the extraction phase, some doubts, about in which cases a metadata should not be translated, arise. The cases which generated a deeper analysis on whether to keep them into consideration or not can be summarised in four categories: not tagged data, wrongly tagged data, non-tagged references and data not identified in the gold standard. For what concerns the first type, the elements not tagged, the decision is straightforward. They are not considered in the translation phase. Indeed, the absence of tagging can be suddenly recognised as missing identification of the content type and, therefore, it is not required an evaluation to check if the metadata are correct. The absence of a tagged metadata can be considered as error even without further analysis, but at the same time it cannot be considered as an error in the evaluation phase since there is no way to identify it, it is simply text not considered to be part of the citation. Regarding the case of badly tagged metadata, e.g., `<lpage>34</fpage>`, the solution is a little more complicated. Indeed, the fact that during the parsing phase the tool tags an element in the wrong way is a complex problem from the point of view of the translation. While on the one hand there is no clear way on how to translate it, on the other hand it should be counted as an error during the evaluation phase. In order to solve this issue a concrete solution is to translate the data with the opening tag but as an empty element. In this way the information won't be lost but it will be counted as a wrong occurrence. In a similar way, in case a reference is entirely composed by a non-tagged string of text, that is translated as an empty reference. In so doing, the reference is counted as a wrongly considered reference, adding a negative point during the tool evaluation. The last issue regards the metadata tagged by the parser but not in the gold standard. This scenario is derived by the fact that in the gold standard only the relevant metadata are considered. While this list cannot be previously applied to the tools, at the same time these metadata should not be identified as errors. Indeed, while on the one hand the metadata present (or not) in the gold standard are known before, we do not know a priori which type of publication is the one of each single parsed reference. This is the reason why all the elements tagged with elements present in the

gold standard for at least one publication type, are translated and will be evaluated later. Instead, the elements which are never present in the gold standard are not translated at all. This difference is necessary since, if a metadata is never considered in the gold standard, it means it is not required to be translated and, also, there is no corresponding element which could be attributed to it. For instance, while the URLs, which are almost never considered relevant metadata apart from the URLs, are translated, while the edition elements, which are never taken into consideration are not considered since there is no way they could be evaluated in comparison to the gold standard. In fact, reporting more metadata than the ones required, as standard metadata for the publication types, does not provide a higher score.

In chapter 4.3 Conversion Script to TEI XML the software, their respective metadata tables list and notes about the features and measures necessary to create the conversion files script are reported.

3.4 Comparison and Evaluation

The final step implemented in this methodology, the evaluation of the tools extraction and parsing performances is based on the combination of two processes. From a chronological point of view, the first action to carry out is the comparison between output file (converted to TEI XML in the previous step) and gold standard. On the other hand, the second challenge to face consists in the evaluation of the results obtained in the comparison phase. However, from a logical point of view, the parameters concerning the evaluation must be decided before the comparison ones. In fact, to understand how to compare the output and the gold standard, it is necessary to know what we are interested in to evaluate. Indeed, on the basis of the parameters selected to evaluate the output quality, it is possible to know at which level to compare it against the gold standard.

Thus, as regards the evaluation phase, there are two questions that deserve to be considered: what to evaluate and how to evaluate. As concerns the former question, in the literature the usual aspect evaluated to define a parser's quality is the number of correctly identified fields for each reference, e.g. (Indrawati, Yoganingrum, and Yuwono 2019; Tkaczyk et al. 2018a). Nonetheless, in order to get a complete evaluation of the tools, the single metadata level may not be enough to show enough features and capabilities. There are other levels of analysis which can be interesting to evaluate to get a complete overview of the tools' potentials. Overall, three levels of analysis have been identified. Each of them represents a different deepness level with respect to the task of extracting and parsing. The three mentioned levels are:

1. **Correctly identified references**. This is the first level among the ones analysed in a chronological perspective. Indeed, a formal evaluation of this aspect provides valid information about the goodness of the tools in the task of identifying the references as single units in the bibliographic section. The focus is on the ability of the software to distinguish each reference, on the one hand, from the surrounding text and, on the other, from the other references. This analysis level provides a perspective on the ability of the tool to carry out the first of the two tasks identified in this research, the reference extraction. Analysing this aspect, concretely, means to identify how many references have been identified by each parser, whether more, less or the same number as the gold standard ones and how many of them have been correctly identified.

2. **Correctly identified fields per reference**. This level goes deeper in the analysis levels. The aim of this type of analysis is to offer a glimpse into the number of correctly tagged metadata have been identified, independently from the content correctness. This aspect is relevant since it allows us to see, from a concrete point of view, the quality of the markers' usage by the tools. In other words, we want to test whether the tools are good in the identification of the metadata, independently from the actual goodness of the text parsing. For instance, this aspect allows one to check if a metadata is inflated in the economy of the tagging distribution, or if, vice versa, it is never used even in cases where it should have been applied.

3. **Correctly identified contents per reference**. This is the deepest and last level analysed. It is the level of analysis commonly used for the definition of the tools' quality. It is basically able to state how many parts of the textual reference have been correctly parsed and identified. This step regards the parsing level par excellence. Specifically, its scope is to check whether the text inside a correctly identified metadata is correct, or at least, sufficiently similar to the gold standard one to be considered correct.

The analysis on these three levels may provide possible scenarios in which, for instance, the same software can be defined good in one of the tasks and bad at the remaining two. Hence, it may happen that a tool is able to identify the references in the bibliography but not their inner data. It is clear that the most influencing level among the ones analysed is the one of the metadata content correctness. For instance, if having to choose between two tools, the first being good at recognising the correct number of references but providing a bad tagging, and the second not being precise with the number of references but providing good tagging and text parsing, this last one would be prone to be accepted. Nonetheless, the perspective of analysing different levels may be concerned with a different final perspective, less related to the identification of a unique tool. Indeed, it may come out that being two

or more tools good at different tasks a combination of them would provide better results than all of them used singularly. Or, vice versa, a same good or bad result in all the three tasks may simply confirm the static excellence or badness of the tool in extracting and parsing the bibliographic section.

As concerns the second question of the evaluation, how to evaluate the results of the extraction and parsing, the answer is more straightforward. Indeed, for the purposes of the analysis at all the three levels defined, the selected parameters are precision, recall and f-score. These parameters have been selected, first of all, since they represent an affirmed type of quality measurement of the results obtained in the information retrieval field. In second place, these measurements are widely used also in the comparison studies similar to the current one. The advantages derived by the use of these parameters are mainly related to the fact that they provide specific measurements on the level of accuracy of the data retrieved. Indeed, the precision computes the ratio of total number of rlevant papers retrieved to total number of retrieved papers. In this way it is shown how precise the software has been in the identification of the references, i.e., how many wrongly identified or missing references are there with respect to the correctly identified ones.

$$precision = \frac{relevant\ items\ retrieved}{retrieved\ items}$$

The recall instead computes the ratio of the total number of relevant items retrieved to the total number of relevant items. This measure shows how many of the correct entries have been retrieved by the extraction tool in the output file, independently from the total number of references identified.

$$recall = \frac{relevant\ items\ retrieved}{relevant\ items}$$

The scope of the f-score (F1), at this point, is to show the tool's general level of accuracy by balancing the values of precision and recall. For this study it has been selected the balanced f-score, representing the harmonic mean of precision and recall.

$$F1 = 2\,x\,\frac{precision \cdot recall}{precision + recall}$$

Even if it is quite a simple method to evaluate the results, nonetheless it is enough to provide a wide view on the topic.

The comparison phase originates from these two aforementioned considerations. Its aim is to compare the parser's output files to the gold standard to get the data through which to evaluate the results. The comparison script implements the two points described before. It must be built in such a way that it

extracts from each file of the output the values necessary to compute the evaluation of all the three analysis levels. Even if the final aim is to compute the total scores among all the files extracted by each parser, in order to get information on all the subparts of the dataset, precision, recall and F1 values are computed also on the single papers and on the single fields. In this way it will be possible to have more information on the single sub-parts and check whether there are specific conditions under which some parsers work better or worse. From a concrete point of view, the process consists in the one-to-one comparison between each single output file converted to TEI XML in the previous step, to its respective gold standard file. The comparison script is based on the fact that all the references retrieved in the output files are considered and translated even if they are more than the ones required in the gold standard. The values computed for each reference of an output file are summed in a unique value which will be summed to the values of all the other files. From that value the final evaluation is computed. The flowchart shows the generic process followed in order to retrieve the necessary information. In the following three sections the specific considerations concerning each of the three layers of the analysis are explained.

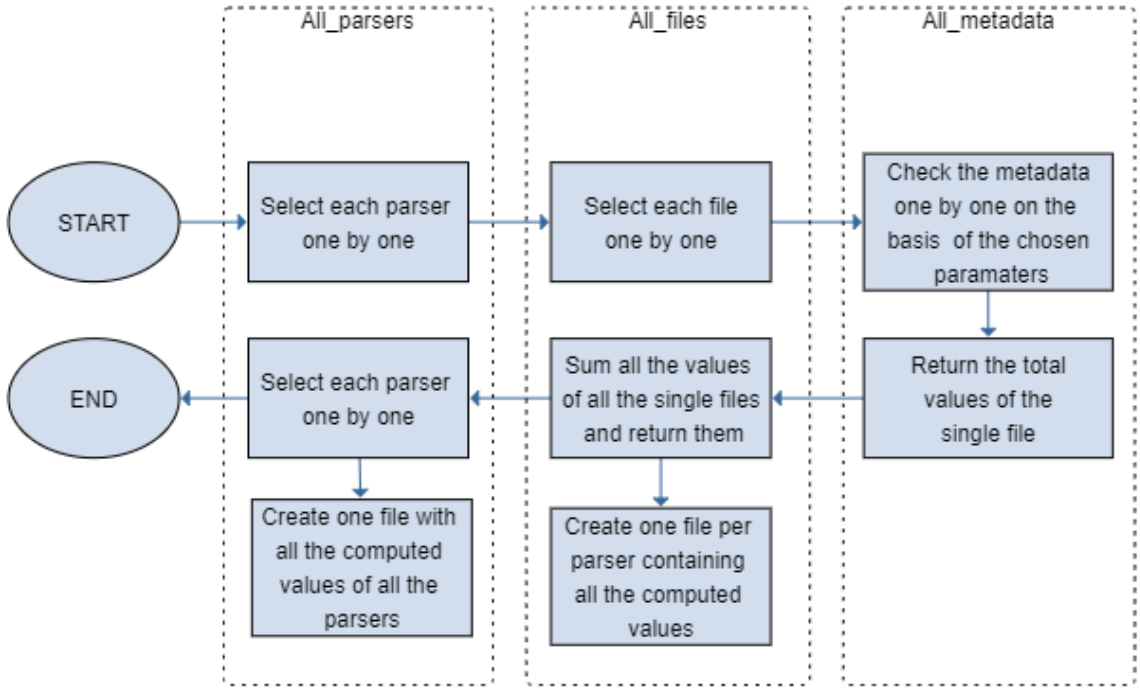

Figure 4. Representation of the workflow of the comparison script.

4. Data

4.1 References Extraction Tools

After the application of the rules stated in the methodology for the selection of the parsers the following seven have been selected: Anystyle, Cermine, GROBID, Excite, Pdfssa4met, Scholarcy and Science Parse. In this chapter they are presented with a theoretical description and a practical guide on how to use them.

The **ANYSTYLE** tool, created by (Keil et al.), is mainly thought to parse single references. Nonetheless the project is provided with a gem able to deal with PDF full texts. For the purpose of this research Anystyle client[29] has been used. This tool is based on machine learning algorithms and there are no further dependences apart from pdftotext[30]. Anystyle is programmed in Ruby. The actual version supported is not explicitly provided, but it is working with Ruby3.1. Anystyle provides the output files in JSON, XML, RIS, BibTeX etc. The output selected for this study is JSON. Indeed, the XML format has some limits among which the lack of distinction between the different authors. The service is available on GitHub[31] as a repository which can be downloaded and used locally; as a website demo[32]. Sadly, the demo is devoted only to the reference string parsing and not to the references extraction and parsing from PDF. Nonetheless, it may be useful for better understanding the service. Anystyle client is provided with accurate documentation[33], which is useful in case of download. The repository's last update, in the current moment, results to have been made on 28th November 2021. The download is pretty straightforward, no specific problems have been found in downloading the tool. But in case something happens the issues page on the main GitHub repository and in the client repository are available with the main doubts (both open and closed).

CERMINE[34] (Content ExtRactor and MINEr) is a Java based tool, aimed to extract information from born-digital only PDF files. The Cermine extraction system uses supervised and unsupervised machine-learning techniques to carry out three major tasks: text extraction, header metadata extraction and references extraction. The references extraction workflow is based on three major steps: the extraction of the paper's layout information, the reference extraction and the reference

[29] https://github.com/inukshuk/anystyle-cli
[30] https://github.com/jalan/pdftotext
[31] https://github.com/inukshuk/anystyle
[32] https://anystyle.io/
[33] https://rubydoc.info/gems/anystyle
[34] https://github.com/CeON/CERMINE

parsing. The former step is based on the analysis of the geometric features of the paper analysed and on their consequent labeling. This step is used as basis for all the extraction tasks carried out by Cermine. The information extracted in this step is used as input for the second one, the references extraction one. Indeed, the tool searches for the part of the text labeled as REFERENCES and, once identified it, it proceeds with the extraction of the single instances with a K-means based system. Finally, through a CRF based process, it parses the extracted references and returns them in the selected structured output format. Cermine supports either HTML or XML Jats as output format. Cermine is available either as an online tool[35], as a web service usable with post requests or as a java library to download and use locally, i.e., a .jar file to call from the terminal. The tool is presented in two main papers, one by (Tkaczyk et al. 2015) and by a slideshow, again provided by (Tkaczyk et al. 2014).

There are a few things to point out about how Cermine works. For instance, the online tool has some problems in identifying the presence of some metadata like URLs and publication places. Moreover, it happens that those elements are tagged with the `<source>` element, which is not a generic tag, but rather a really specific one[36]. Indeed, the scope of this tag is to identify the source in which the work cited is contained. It is valid both for the journal where an article is published or the book which contains the cited chapter. Thus, all the occurrences of this tag with other contents should be translated with the meaning of `<source>` and then, considered as an error once compared to the gold standard. The publishers and the publication places are not recognised as such and usually tagged with other elements. In the same way URLs, DOIs and digital pages are not recognised nor tagged.

EXCITE (Extraction of Citations from PDF Documents) is a project including a number of citation extraction software. In particular, the software which directly regards the references extraction is called ExParser. This tool has the ability to extract and parse the references in XML. This is why the ExCite process is based on a three steps workflow. The first step consists in the extraction of the input PDF paper layout information, and it is carried out by CERMINE. Then, the layout, together with the text in XML returned as output by Cermine, is used as input for the ExParser tool which extracts the references and parses them. Possibly, ExCite provides, also, a tool for reference matching, EXMatcher, but this is out of this research's necessities. ExCite is available either as a repository to download and as an online demo. In both the tools, the process followed for the references extraction is the same. While the online demo is straightforward to use, downloading it locally means some more work. Indeed, the tool must be trained before being used. The tests through which train Exparser

[35] http://cermine.ceon.pl/cermine/index.html
[36] http://jats.nlm.nih.gov/archiving/tag-library/1.1/element/source.html

are provided in the project repository but for a non-expert user this is a more difficult step in order to use the tool.

GROBID, is a CRF based tool for extracting and parsing text in unstructured format. Its objective is to return the text extracted, for instance from PDF papers, into structured XML/TEI format. The GROBID references extraction workflow is based on three steps, through a "cascade of sequence labeling models".[37] Indeed, first of all, the full text is segmented and labelled line by line. Then the references section is identified and further segmented in the single citations it is composed of. Finally, each citation is segmented in its subparts and converted to TEI XML. Through this process, apart from the references, GROBID is able to extract all the various sections of which the articles are composed of. In particular, it is widely known for its quality in the header extraction task. GROBID is a service available either as a website[38] or as a library to download.[39] In both cases by implementing the server it opens an interface through which it is possible to load the pdf files and get them back parsed and structured as XML TEI documents. No particular problems have been registered in the download phase nor in the local usage.

PDFSSA4MET is a tool that makes use of regular expressions in order to extract the references from files in PDF format. Its usage is dependent on the presence of the software pdf2xml. It is basically used to extract the text from the PDF to the XML language, and from that file works pdfssa4met. It is written in Python, in a version of the programming language going from 2.6 on. In any case, since its last update dates back to 2013 it doesn't work with Python 3, since it is used the syntax for Python 2 (for this work it has been used Python 2.7). Pdfssa4met is available on GitHub[40]. An interesting aspect is that thistool can extract the paper's metadata, title and authors. Pdfssa4met can be used either by command line or in a Python IDE by typing in the main call of the file "references.py" the path to the PDF file to parse. In both the cases the path to the pdf2xml library must be updated manually in the code. Also, in order to use Pdfssa4met, it is necessary to register to the OpenCalais Web Service in order to get an API token that will need to be added to the `config.py` file.[41]

[37] https://grobid.readthedocs.io/en/latest/Principles/
[38] https://cloud.science-miner.com/grobid/
[39] https://grobid.readthedocs.io/en/latest/
[40] https://github.com/eliask/pdfssa4met
[41] In "config.py" it is said: "You'll need to register for one yourself at http://www.opencalais.com/user/register, in order to get an API Key for OpenCalais Web Service" actually that site is no more available. I found a useful page at *https://developers.refinitiv.com/en/api-catalog/open-perm-id/intelligent-tagging-restful-api/quick-start* and then I registered at *https://my.refinitiv.com/content/mytr/en/register.html?_ga=2.78691037.877694478.1637676184-1213138658.1637676184.* After getting registered, it is to create the password (in order to create it you have to wait for an email which will allow to create it), then access the website *permid.org*, log in with the created credentials, select the button "APIs" and finally select "display my API token" which will show your 32 character token. Add it to the config.py.

There are some problematic aspects to point out. The software has problems in case the title or one of the chapters titles includes the word "reference(s)". For instance, when testing the tool on the paper 'RefUTU: Automatic Bibliography Database Generation for Freely Formatted Reference Listings' (Holvitie and Leppänen 2015), it considers as references all the titles, text and page numbers as references (and the actual references are completely ignored, maybe because the word reference has already been retrieved and other instances are ignored). Also, two errors that happened and that made it difficult to approach the tool have been the error: "`sh: /usr/local/bin/pdftoxml.exe: Permission denied \n Could not convert to XML: [Errno 2] No such file or directory'`. To make it work I needed to get the permission to execute it, by using the command "`chmod 755 PATH-OF-THE-FILE`".

Scholarcy is a tool which makes use of an API for extracting metadata from PDF and docx files. While it is able to extract only the references, it accepts as input, apart from the files in .pdf, the .docx and text files. The files can be either posted from the local machine or got with a URL. Scholarcy supports BibTeX, CROCI, Jats, RIS, text and XML as output formats. It is available as an online open source.[42] The literature is scarce, but a paper can be found (Gooch 2021). There are two comments to make on how to use the Scholarcy API interface. First of all, the output in Jats is more accurate than the one in XML, in terms of precision of the data identification. Then, for the online tool it is better to select **v2**, which is more precise in identifying the Jats metadata. This does not apply to the references which, instead, are retrieved in the same number as v1. To conclude, the online API interface supports only files of small dimensions, thus heavy files are not parsed, and a server error message is returned as response.

Science Parse is a tool based on a heuristic labelling process. Science Parse depends on PDFBox for the extraction phase, from PDF to tokens containing paper's layout information. These tokens are used as input for an LSTM model which finds the strings in this model and extracts them. It can extract title, author(s), abstract, sections and in-text citations. Science Parse is a library and an online tool based on Java. The Java version required to work with it is 11.0.12 or less, (even jdk 8 is fine). Even if Science Parse is maintained, some problems have occurred when trying to use it. First of all, it was almost mandatory to use science parse version 1 in its version 2.0.3. Indeed, a version 2 was created but results are currently not updated and not usable. Also, version 3.0.0 of the project's first version, presented problems in downloading the required Scala libraries, which resulted in it being corrupted. Also, another difficulty of using Science Parse is the fact that in order to use it, it is required

[42] https://ref.scholarcy.com/api/

at least a 6GB RAM and a Windows operating system. Thus, the best option for using Science Parse is downloading the jar file for the client, '*science-parse-cli-assembly-2.0.3.jar*' on a computer with the specified system requirements.

4.2 Gold Standard

For the creation of the gold standard all the tasks identified in the methodology have been carried out. For what concerns the selection of the papers, the first instruction was to select a wide variety of publication fields. To test the goodness of the extraction tools against a wide variety of publications, the gold standard is based on 56 articles, written in English, selected from 27 different academic fields (from the humanities to the social sciences and the scientific fields), see *Table 2* to check all the fields. The publications used as a basis for the gold standard have been selected from the resource (Santos, Peroni and Mucheroni 2022). This report has been selected as a source of information as it is based, in turn, on an accurate study on the different typologies of publication. Therefore, to get a wide and accurate selection of articles to pass as gold standard for the software testing, two articles for each topic have been selected, the first two in logical order for each folder (Cioffi 2022b).

The articles have been randomly selected between the papers which, for each field, presented two characteristics: presenting an explicitly mentioned references section and being a non-scanned document. For what concerns the former, to be accepted a paper must be provided with an explicitly mentioned "References" or "Citations" section. Overall, 54 out of 56 papers, of which the testing subset is composed, have been selected on this basis. Nonetheless, in order to provide a wider view on the tools extraction capabilities, two papers were inserted, one not containing at all a References section, and the other containing a section for the references but called in another way. In this way it is possible to test the efficacy of the tools against non-conventionally structured articles. In the end, there will be an evaluation specific for these two papers which will provide specific insights into the tools capabilities to identify the references even if no "References" label is set.

Table 2. Research fields, related short names, papers DOI and short papers names

RESEARCH FIELD	SHORT NAME	PAPERS DOI/ REFERENCE	PAPERS NAMES
Agricolture and Biological Sciences	AGR-BIO-SCI	10.1017/S1751731119000570	AGR-BIO-SCI_1
		10.1007/s13197-019-03938-9	AGR-BIO-SCI_2
Arts and Humanities	ART-HUM	10.1525/mp.2019.37.1.66	ART-HUM_3
		10.1080/03057240.2019.1573724	ART-HUM_4
Biochemistry Genetics and Molecular Biology	BIO-GEN-MOL	10.33594/000000168	BIO-GEN-MOL_5
		10.1016/j.pep.2019.05.004	BIO-GEN-MOL_6
Business Management Accounting	BUS-MAN-ACC	10.1108/IJCHM-10-2018-0849	BUS-MAN-ACC_7
		10.1111/1748-8583.12232	BUS-MAN-ACC_8
Chemical Engineering;	CHE-ENG	Journal of Advanced Research in Fluid Mechanics and Thermal Sciences 62, Issue 1 (2019) 43-52	CHE-ENG_9
		10.1039/c9cy01398a	CHE-ENG_10
Chemistry	CHEM	10.1016/j.elecom.2019.106537	CHEM_11
		10.1039/c9dt02938a	CHEM_12
Computer Science	COM-SCI	10.1007/s11554-017-0669-4	COM-SCI_13
		10.1109/LCA.2019.2935445	COM-SCI_14
Decision Sciences	DEC-SCI	10.1007/s00291-019-00553-0	DEC-SCI_15
		10.1016/j.datak.2019.101721	DEC-SCI_16
Dentistry	DEN	10.1111/clr.13514	DEN_17
		10.1002/JPER.19-0049	DEN_18
Earth and Planetary Sciences	EAR-PLA-SCI	10.1130/G46491.1	EAR-PLA-SCI_19
		10.1016/j.gca.2019.07.021	EAR-PLA-SCI_20
Economics Econometrics Finance	ECO-ECO-FIN	10.1257/aer.20181897	ECO-ECO-FIN_21
		10.1016/j.frl.2018.09.009	ECO-ECO-FIN_22
Energy	ENE	10.1016/j.rser.2019.109298	ENE_23
		10.1080/00295639.2019.1604048	ENE_24
Engineering	ENG	10.23940/ijpe.19.10.p1.25632569	ENG_25
		10.1016/j.ifacol.2019.11.007	ENG_26
Environmental Sciences	ENV-SCI	10.1016/j.jhydrol.2019.123973	ENV-SCI_27
		10.1080/1523908X.2019.1670048	ENV-SCI_28
Health Professions	HEA-PRO	10.1016/j.jcm.2018.11.005	HEA-PRO_29

		J. Phys. Ther. Sci. 31: 771–775, 2019	HEA-PRO_30
Immunology Microbiology	IMM-MIC	10.1099/mic.0.000835	IMM-MIC_31
		10.1016/j.virol.2019.08.005	IMM-MIC_32
Materials science	MAT-SCI	10.1016/j.apsusc.2019.06.253	MAT-SCI_33
		10.1007/s10570-019-02664-x	MAT-SCI_34
Mathematics	MATH	10.1080/03081087.2018.1481357	MATH_35
		10.1016/j.aml.2019.05.016	MATH_36
Medicine	MED	10.2147/OPTH.S217736	MED_37
		10.1001/jamainternmed.2019.2407	MED_38
Multidisciplinary	MUL	10.1126/science.aaw8848	MUL_39
		10.1038/s41598-019-50584-4	MUL_40
Neuroscience	NEU	10.3389/fncel.2019.00448	NEU_41
		10.1016/j.biopsych.2019.02.010	NEU_42
Nursing	NUR	10.1177/0969733018774828	NUR_43
		10.1111/jonm.12826	NUR_44
Pharmacology Toxicology Pharmaceutics	PHA-TOX-PHA	10.1016/j.jconrel.2019.08.007	PHA-TOX-PHA_45
		10.3390/md17100588	PHA-TOX-PHA_46
Physics and Astronomy	PHY-AST	10.1007/JHEP10(2019)213	PHY-AST_47
		10.1364/OE.27.032378	PHY-AST_48
Psychology	PSY	10.1111/jopy.12444	PSY_49
		10.1037/apl0000399	PSY_50
Social Sciences	SOC-SCI	10.17645/up.v4i3.2210	SOC-SCI_51
		The Western Journal of Black Studies, Vol. 42, No. 3 & 4, 2018	SOC-SCI_52
Veterinary	VET	10.1016/j.jevs.2019.102796	VET_53
		10.1177/1098612X18810867	VET_54
Papers with no “References” label	Z-NOTES-TESTS	10.1177/0008125619849443	Z-NOTES-TESTS_1
		10.1002/wsb.993	Z-NOTES-TESTS_2

Differently, for what concerns the scanned documents, this kind of documents could not be considered since the tools are designed in order to work with digitally created sources. Thus, a scanned document cannot be parsed by such tools. Moreover, this papers’ subset presents a wide variety of external features: more than 1000 different journals and almost 150 publishers (including

books in all their shapes, proceedings and grey literature). The dataset presents an average of 45 reference strings per paper (computed as the mathematical mean of all the reference numbers in all the papers divided by the total number of papers), with a recorded maximum of 113 and a minimum of 10 bibliographic references per paper. The dataset presents 2538 total references, 65 of which in the two papers not explicitly provided with a References section. All the papers in the dataset have been published in 2019, with only two exceptions, the former published in 2018 and the latter in 2020.

Figure 5 shows the distribution of the cited works types in the input datset. In particular, it is to note the discrepancy between the journal articles ('articles') and all the other publications. Since it was not possible to show the comparison of the articles with all the other publications, because of the excessive distance in numbers, two graphics were produced. *Figure 5* shows the comparison between the three most represented publication types. Instead, *Figure 6* shows the comparison between all the publication types excepted by the journal articles. Finally, as another matter of difference, while the papers are written in English, the standard language for reference extractors, some of the works cited in these articles are in a different language. These are mainly German (the most frequent), Dutch and French. Rarely, other languages like Chinese and different African idioms are recorded. Testing the tools against a small subset of non-English references can be considered as a mirror of the effects of language differences on the parsing quality.

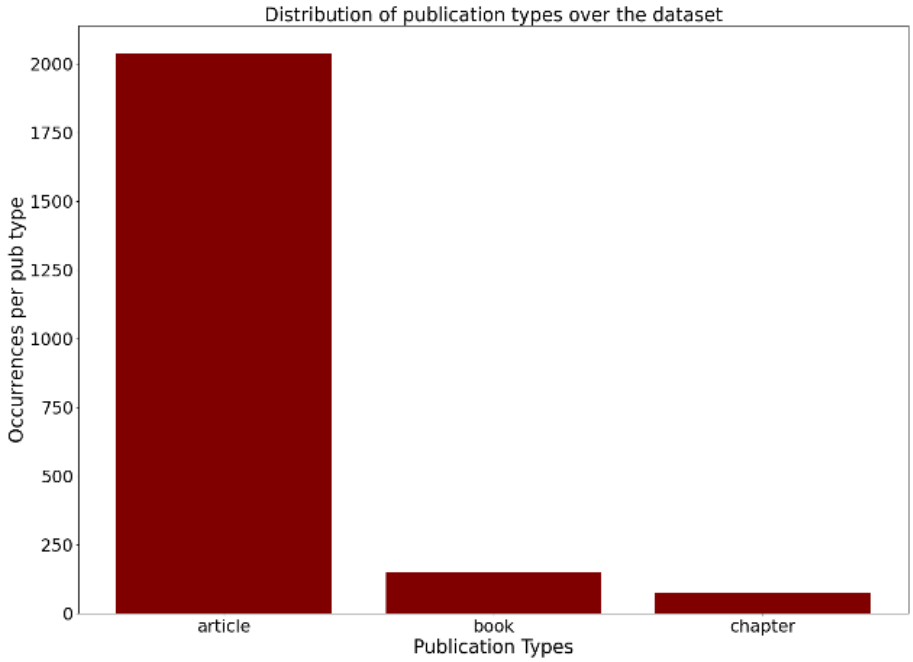

Figure 5. Most cited publication types in the dataset.

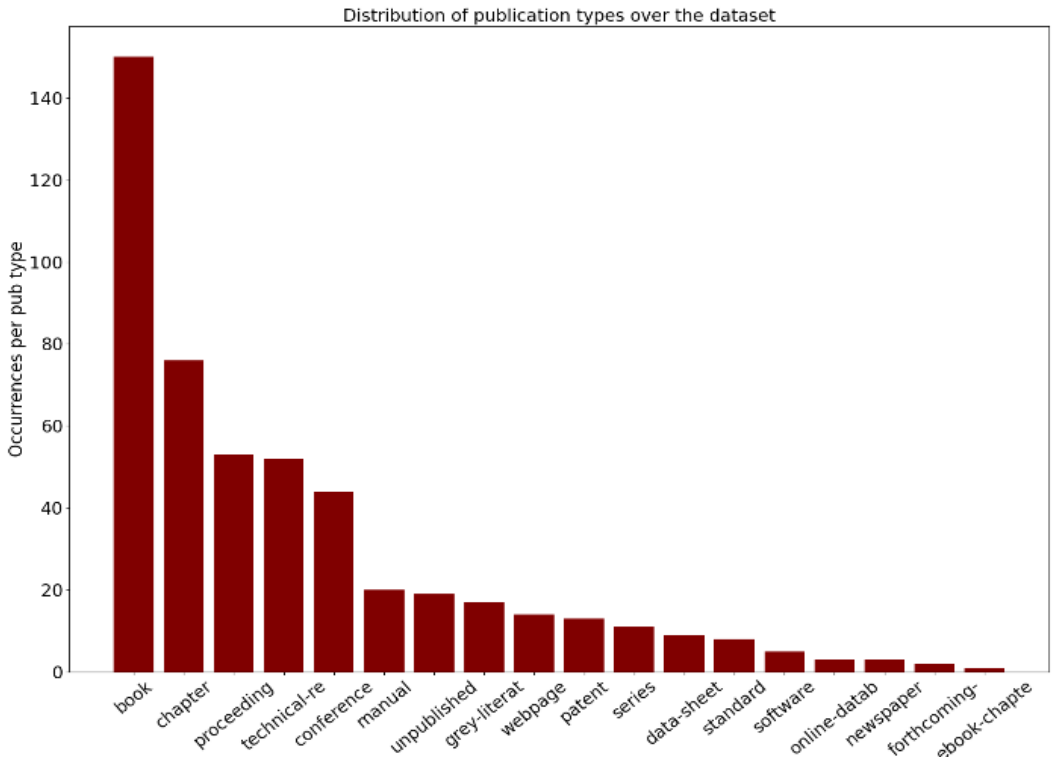

Figure 6. Distribution of the publication types without journal articles

For what regards the second task to carry out, the selection of the metadata to consider, some specific aspects have been investigated to understand how to manage the specific shade of this task. The first aspect to consider, is that the Latin formula "In" has not been taken into consideration as standard metadata. Indeed, its value, that of meta-information describing in which book or proceedings a chapter can be found, is not necessary in a structured form like the one of TEI XML. The same information, indeed, is expressed in the final format through a bibliographic structure based on the specifications of the tag `<title level="">`. Indeed, the article whose title is included in the tag `<title level="a">`, located in the section `<analytic>`, is part of the monograph whose title is expressed in the section `<title level="m">` inside the section `<monogr>`. Therefore, there is no structural need for the "In" formula to be considered in the gold standard XML files.

The second aspect on which the creation of the gold standard was based, is the fact that, as a rule, the references should be interpreted literally. Indeed, the manual references extraction could be performed two different ways: reconstructive and literal. The first one implies the fact that in case of formulas, grammatical or formal errors etc., the text should be reconstructed to have the complete and correct version of the original reference. This version implies that only the software which makes use of external databases can obtain high scores since it is hard to obtain this information simply considering the text itself. The second one, instead, implies that these issues should be left as they

are, and the evaluation of the automatic extraction performed by the automated tools should be based on a standard output. Some of these issues and the respective solution are reported here above:

- the formula “et al.” or the three dots (“...”), used to avoid the explicit mention of all the authors of a publication, is not considered: only the authors reported in the reference string are taken into consideration.
- in case it is provided only the initial of the forename(s), the initial is enough, and it is not searched the entire name (which requires a research on external databases or in the original articles webpage);
- if there are errors (e.g., duplicates, spelling or font errors, copresence of different citation styles), these should be kept. In case some references are particularly rich of errors this aspect should be kept into consideration when making the conclusions out of the results evaluation.

Finally, the third task, the selection of the language of the gold standard was based on two major factors: the first concerning the quality of the language, its possibility of expression and the compatibility with software libraries; the latter, instead, was a more concrete parameter, based on the features of the selected parsers. For what concerns the language ease of usage, the choice was almost easy with the XML (eXtensible Markup Language) meta-language. Indeed, derived from SGML, XML, thanks to its inner extensibility, can describe the texts at different depth levels. Therefore, all the differing shades of publication metadata could be correctly represented. Also, it is a widely used meta-language and, therefore, there are plenty of libraries, in particular for Python in this case, and utilities which can be used in order to read and write the files written in XML. This aspect is relevant in this case, since it provides an easier way to analyse in depth the document structure and find the aspects relevant for the comparison between the gold standard paper and the one produced by the extraction tool. On the other hand, from a concrete point of view, there is another consideration to make. Most of the retrieved software (GROBID, Cermine, Excite, Scholarcy, Pdfssa4met) produce their output in XML, therefore it is easier to create and surer to use a conversion from XML to XML. Indeed, there is a minor risk that some aspects are lost and at the same time it is easier to convert an XML file with certain guidelines to an XML file based on different ones then converting a BIBTEX into an XML.

Once selected XML as the language for the gold standard, the second necessity was to identify the XML standard to use, i.e., the set of rules to follow in order to create the final documents. Indeed, since XML is an extensible language, to avoid the excessive proliferation of different markup elements and, at the same time, to create some reasoned rules, which should become common to the community using XML, various set rules were created and are still used. In this case the choice was,

again, quite straightforward: the TEI (Text Encoding Initiative) standard guidelines. The TEI, in fact, on the one hand provides to the users' guidelines specifically conceived for texts in digital format and, on the other hand, it is widely used and accepted in the academic and library fields. Finally, another practical aspect which has been considered in order to select TEI as encoding standard, is that one of the parsers selected in the research used this encoding as standard for its output files. Even if this last element was not one of the main reasons for which the TEI encoding was selected, it is surely another sign that TEI is widely spread in the digital texts field.

The structure of the XML file and of the single citations is the same for all the 56 articles. In order to create them, the guidelines provided by TEI have been followed, and at the same time it has been kept into account the usage of the metadata made by GROBID.

The first element in the documents is the declaration. It is specified the XML version used and the encoding through the tag `<?xml version="1.0" encoding="UTF-8"?>`. Then, there are two elements, which are "available for the representation of the outermost structure of a TEI document"[43]. Indeed, they define the structure of the document containing the encoded references but do not express the type of text contained in the document. These tags are[44]: TEI and text. The first one, from a dependency perspective, is the TEI element. It is aimed at identifying the presence of a "single TEI-conformant document"[45]. In this tag are conveyed the information concerning, on the one hand, the structural data of the file, including the XML Namespace (`xmlns="http://www.tei-c.org/ns/1.0"`) and, on the other hand, the metadata of the document contained, including the fact that the language selected is English (`xml:lang="en"`). Second and last, the extracted references are contained in a block called `<text>`. Since the bibliographic references are considered as an independent block of information and not as part of the text which they belong to, no other subclass, e.g., body or back, has been used.

Then, the first element which defines the type of text contained in the document is `<listBibl>`. This element is used to define the presence of a list of various references, which will be more specifically defined in tags nested inside this one. This is the last level before the citations are considered singularly.

[43] https://www.tei-c.org/Vault/P5/2.4.0/doc/tei-p5-doc/fr/html/DS.html

[44] Link to the page describing the structure of the standard TEI document: https://tei-c.org/release/doc/tei-p5-doc/en/html/DS.html

[45] https://tei-c.org/release/doc/tei-p5-doc/en/html/DS.html

Inside the <listBibl> element, the references are considered singularly and structured following the procedure described in the TEI website page dedicated to the tagging description.[46] The TEI rules accept two types of expression for the references, one structured and the other unstructured. For the scope of this research, it was necessary to use the structured one since in most cases, with only some exceptions, the references in natural language convey a hierarchical meaning which must be recognised in the XML encoding (e.g., an article and a journal are not two different elements on the same level but they are rather on two different levels, the contained and the container respectively).

<biblStruct> is the references container element. It defines a structured bibliographic entry, and it is provided with an id, identifying each specific reference, and with a type, specifying the publication typology. The id, defined as a key-value pair, is introduced by the fixed key "xml:id". The value, instead, is a variable string composed of 2 parts: the letter "b" which stands for "Bibliography" and the integer of the current reference position in the extracted references list (in a range from 0 to the length of the list minus one). The type is specified again with a key value pair, where the key is "type=" and the key is a string describing the publication type. The string is selected between these values: "article", "book", "chapter", "series", "conference", "data-sheet", "ebook", "ebook-chapter", "forthcoming-article", "grey-literature", "manual", "newspaper", "online-database", "patent", "preprint", "proceeding", "software", "standard", "technical-report", "technical-report-chapter", "unpublished", "webpage". Under the 'manual' voice fall manuals, toolkits and guides, as grouped in the original article. In case any of the cited works did not precisely fit this classification, it was associated to the one which seemed the most similar. It is the case of the thesis typology, which was considered together with the grey literature. Indeed, the thesis does not have a precise definition and is historically associated with grey literature. Also, in the same article the grey literature is defined as a possible wider group for the thesis type. Differently, the protocols and the laws were absorbed by the standards class. Indeed, that was the publication type with the most similar definition and number of fields identified. The workshops and the symposia are considered under the definition of conference.

<analytic>, <monogr> and <series> are the three inner structures contained by the <biblStruct> element. These elements, in their turn, contain the actual reference metadata. The <analytic> element is used for the information concerning "item [...] published within a monograph or journal and not as an independent publication".[47] For instance, the analytic section will contain the information regarding a journal article, while the article information will be inserted inside

[46] https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html
[47] ibidem

the <monogr> element. This second element, in fact, carries the information of independent publications, e.g., books and journals. Finally, the series element contains the specifications of the series which an item belongs to.

For what concerns the concrete usage of the encoding, the official TEI guidelines page[48] have been followed. Here some details are provided:

- The Greek alphabet letters, the special characters, the symbols and some accented characters have been translated in the corresponding XML accepted character, following the UNICODE specifications.[49]
- when more than one option was available to define a specific content type, it was selected the one used in the GROBID encoding. This choice is merely pragmatic, due to the necessity of comparing the two types of XML files

In Appendix A - Metadata Tagging and Usage it is shown the final version of the encoding selected for each type of metadata, together with its logical position in the encoding structure, identifiable in (Cioffi 2022b).

4.3 Conversion Script to TEI XML

The evaluation scripts have been created on the basis of the specific necessities of each tool and language. As concerns this research, whenever possible the XML format was selected as language for the output file for a matter of syntax complexity and automatic manageability. Indeed, Scholarcy and CERMINE are provided with Jats XML, option selected because of its formality; Pdfssa4met and ExCite are outputted in generic XML, without specific rules; finally, Science Parse and Anystyle output is in JSON format. With a mixture of these elements and the requirements of the language the scripts were created. They are presented here below one by one.

The steps required to proceed with the creation of the conversion script are the following ones:

1. Create a flowchart to understand how to develop the steps required to carry out the conversion script.
2. Find a Python library to use to create the script (LXML and Etree in this case).

[48] https://tei-c.org/release/doc/tei-p5-doc/en/html/CO.html
[49] http://www.unicode.org/charts/

3. For each file create the script code of conversion (in case some tools tags are close, use the same file).
4. At the end three file types for each tool will be created: the original output file (Cioffi 2022b), the script file (Cioffi 2022a) and the output TEI XML file (Cioffi 2022b).

For what concerns the first tool, Anystyle, the output is provided in JSON format. Thus, the process must be a full conversion to XML TEI. Anystyle has the widest used set of metadata among all the retrieved tools. Indeed, it is the only one, together with GROBID, which provides the metadata necessary to explicitly cover all the fields used in the gold standard. It also provides more metadata than the ones required and applies the publication type recognition. The value corresponding to the key "type", which is not translated to TEI since it is not a required information, reports the publication type identified by Anystyle. This instrument is particularly useful to identify the presence of the formal sections to translate: `<analytic>`, `<monogr>` and `<series>`. The case study is reported in Table 3. Anystyle metadata and TEI conversion. Among the selected metadata the only problematic aspects regard the lack of identification of starting and ending pages. Indeed, both the page ranges and the single pages are identified through the key "pages". The authors and editors are meticulously classified: reported as a list of dictionaries, they can be described by a set of metadata: "given", "surname", "particle" (dedicated to the nominal particles like the italian "de" or the Dutch "van"), "literal" and "others". This last one is a particular case because its expected value is a Boolean (true) present in case the formal structure "et al." is reported at the end of the authors list. Instead, the metadata "arXiv", identifying the arXiv id, "citation number", identifying the possibly present citation number in the bibliographic list, "edition", "others", "source", "translator" and "type" are not reported in the converted file.

Table 4. Anystyle: analytic, monographic and series, instead, shows the cases in which each section should be created in the references. While the monographic section is always present, the analytic section is inserted only in case the "Type" node has one of the values specified in the second cell of the first row. Finally, the book series are not signaled in the "type" values as such, but they are, instead, specified in an inner tag inside the reference, with the specific key "collection title."

Table 3. Anystyle metadata and TEI conversion

Json	TEI XML
“arXiv”	-
“author”	<author>
“citation-number”	-
“collection-title”	<series>
“container-title”	<title level=”[m | j]”>
“date”	<year>
“doi”	<idno type=”DOI”>
“edition”	-
“editor”	<editor>
“family”	<person> <surname>
“genre”	<note>
“given”	<person> <forename>
“issue”	<biblScope unit=”issue”>
“literal”	[text of author tag]
“location”	<pubplace>
“note”	<note>
“others”	-
“pages”	<biblScope unit=”page”>
“particle”	<surname>
“publisher”	<publisher>
“source”	-
“title”	<title level=”a”>
“translator”	-
“type”	-
“url”	<ref src=””>
“volume”	<biblScope unit=”volume”>

Table 4. Anystyle: analytic, monographic and series

	"Type": ["chapter" | "article-journal" | "paper-conference"]	"Type": ["report" | "book" | null]	"collection-title"
analytic	X		
monograph	X	X	X
series			X

For what concerns Cermine some specific aspects were taken into consideration in order to carry out the conversion. As seen in the previous chapter, Cermine provides as output an XML file in Jats format. The fact that it uses Jats, on the one hand, is useful, since it provides a set of tags which should be used to produce a good document. Also, one automatic tool has been created to automatically transform the Jats format to TEI[50]. Nonetheless, Jats provides a really huge set of metadata and its usage provides an even wider number of combination possibilities which makes it difficult for an automatic conversion for all the parsers outputs. Also, the conversion should be guided to a specific standard, that of the shape provided to the gold standard. Because of these reasons, the option of the automatic parser was rejected, and specific parsers for each XML Jats output was created.

One thing to notice is that Cermine does not present a wide variety of tags. *Table 5.* Cermine metadata and TEI conversion shows all the tags retrieved. They do not allow for a complete recognition of the metadata: the analytic and monographic section should be deduced by pragmatic aspects: `<article-title>` and `<source>`. The series section, the issue and the URIs are not recognised. Vice versa the names, years, volume and pages are fully recognised.

It is noteworthy the case in which two or more nodes with the same tag occur in the same reference. The only case in which can be identified as correct, and not a duplication of a badly identified tag, is the tag `<string-name>`. In all the other cases this must be considered as an error and handled consequently. Indeed, there is no case where two dates or two article titles have been reported in the same bibliographic entry In case two dates are identified as the beginning and the end of a publication process, these should be represented as one single unit separated by "-". In this work, only the first of those elements is considered and translated in the new file. Even if this can be considered as a strict reading, it was essential to take a decision required by the concrete circumstances. Indeed, by looking at the results of Cermine, it is clear that some tags are used in the wrong way. Therefore, there was

[50] https://github.com/kermitt2/Pub2TEI

the necessity to decide how to handle those exceptions. The decision to consider only the first occurrence of the metadata was taken considering the necessities of the final comparison. For instance, the presence of a <source> tag is relevant to define whether the monograph section should be created or not. Indeed, the basic principle of the division between analytic and monographic sections is that, whether there is a work (identified with the tag <article-title>) and a container of the work (identified with the tag <source>[51]), the two aspects are separated in these two categories. Also, the definition of the monographic and analytic sections is relevant since they define the concept expressed by 'In' in the references. Thus, unless it is impossible to identify the two sections, they should be recreated through the metadata reported by the tools which allows to recreate this structure. Nonetheless, <source> is used in wrong ways by CERMINE and in many cases <article-title> refers to books (and not to book's chapters), so this passage is difficult to define in both a conceptual and a formal way. In particular, it is hard to identify a book chapter where both the chapter and the book are identified as <article-title>. There were different options on how to handle this case, from not recreating the two sections and just checking if one of the <source> occurrences was the correct one to simply considering as wrong all the occurrences of <source>. Ultimately, the one which seemed the most significant, in order to maintain both the meaning carried by the macro-structure "analytic/ monographic" and by <source>, is just considering as correct the first occurrence and on its basis create the monographic section. At the same time, the contemporary presence of two <article-title> elements does not allow to identify them as chapter and book, since the meaning of the element should be completely changed in function of a wider acceptance. Thus, it is already clear since this step that the book chapters will be rarely identified by this tool.

Table 5. Cermine metadata and TEI conversion lists in the left column the metadata used in CERMINE in order to define the citational data. In the right column there are the respective translations in XML TEI. All the elements not reported between a JATS element are not translated since they can be considered as non-tagged elements, and therefore not recognised. This table is based on the elements actually retrieved in the CERMINE results rather than on the JATS rules since there is such a freedom in NLM syntax that it is hard to only theoretically define the usage made of it in the current tools. Just to make an example, the author should be defined by the syntax "<person-group person-group-type="author">", but CERMINE identifies the authors with the generic element "<string-name>". Thus, it is impossible to distinguish between authors and editors. This is an aspect that will be handled in the comparison phase. The syntax is rather used in

[51] http://jats.nlm.nih.gov/archiving/tag-library/1.1/element/source.html

order to verify whether the software correctly uses all the elements. The metadata which are not identified by Cermine are: DOI, note, publisher, publication place and URL.

Table 5. Cermine metadata and TEI conversion

NML/ Jats (CERMINE)	TEI
<article-title>	<title level=”a”>
<fpage>	<biblScope unit=”page” from=””>
<given-names>	<forename>
<issue>	<biblScope unit=”issue”>
<lpage>	<biblScope unit=”page” to=””>
<mixed-citation>	/
<ref id="ref[n]">	<biblStruct id="b[n-1]">
<string-name>	<author><persname>
<source>	<monograph><title>
<surname>	<surname>
<volume>	<biblScope unit=”volume”>
<year>	<date when=””>

Table 6 shows the cases in which the analytic and the monographic section should be kept in the converted TEI XML file. The analytic section is present only in case there is a work and the work container (<article-title> and <source>) or the work container (even if no work is explicitly mentioned). The second case, in particular, is voted to recognise the cases in which the journal title is reported without the article title in the references, which happens systematically in 2 out of 56 papers, and rarely in other papers. In case only <article-title> is present, it can be considered that there is no actual article, since there is no case in which the article is present without its journal. It should be rather considered that the <article-title> metadata has been wrongly attributed to a book or report title. Finally, if none of them is found, then a section monograph should be created anyway since in none of the gold standard citations there is a citation without, at least, a monographic section (and, thus, a title). Even if, from a concrete point of view, the element <article-title> is used for articles even without the <source> element (which usually is wrongly considered as a separate citation, and possibly defined as <article-title> too), article title is also used to identify monographic resources. Also, from a theoretical point of view, it is not

possible to identify an article without its journal or a chapter without its book because a huge piece of its identification would be missing, and, from a practical point of view, this scenario has never been registered in the dataset while the opposite is found, i.e. it is provided the journal title without the article title.

Table 6. Cermine: analytic and monographic

	+ <article-title> + <source>	+ <article-title> - <source>	- <article-title> + <source>	- <article-title> - <source>
analytic	X		X	
monograph	X	X	X	X

Differently, Excite is a tool available to the final user as an online interface providing the output files in either XML or BibTeX. For the purposes of this research the XML meta-language was selected as the preferred one. Indeed, by looking at the accuracy of the recognition, and to the metadata nodes typology, it was clearly an optimal solution for the purposes of this project. The conversion code has been created in union with the Pdfssa4met output files. Indeed, even if ExCite has a higher level of metadata recognition than Pdfssa4met, they share some common metadata. Also, since the structure of the conversion file allows for a wide metadata recognition, there are no restriction problems for a wider range of metadata typologies. Nonetheless, since the metadata are outputted in a generic XML format, without any set of rules being specified, the metadata have been manually identified from a set of ten testing papers. *Table 7*, shown below, has been created by a methodic empirical approach.

One negative aspect of the website to point out is that the download is not currently possible. In order to get, nonetheless, an output XML file, the references have been manually copy-pasted in an XML file, included between the following structural metadata (Cioffi 2022b):

- XML declaration,
- <references> section,
- <reference> section (one for each reference identified by ExCite).

Table 7 provides the potentially identifiable metadata. The range of identifiable metadata is pretty wide, in line with GROBID, Cermine and Scholarcy, from a quantitative point of view. In this case, there are two types of metadata which, even if existing, are ignored in the translation. These are <edition> and <other>, normally used for publication place or information not identifiable with other metadata. Both of them, in fact, are not consistent with respect to the standard metadata

considered in the gold standard. The <issue> tag is translated but we can already state that it won't probably be considered by the system. This metadata indeed is reported only in case the referenced item is a book series. Since in none of the tested papers it has been reported an element containing the series-title information, it is supposed that the series are not identified. In this tool it is preserved the difference between the first and the last pages of the cited work. Also, since the tag <source> does not allow for a deeper identification of the publication typology, the title in the monographic section is not further identified with the attribute <title level="[j | m]">.

Differently, the article title is defined with the attribute <title level="a"> if a <source> element is reported, otherwise only as <title> located in the monographic section. Indeed, just like for Cermine, it has been recorded that in case an entire book, and not a single book chapter is mentioned, it is identified with the <title> tag. This is different in case only the journal title is reported without the respective article. In that case instead, it has been reported that in all the cases the journal is identified as such, with the tag <source>. Thus, in case only the article title is reported it is created a monographic section. Instead, if only the source record is found, then it is translated as <analytic> plus <monogr>. Finally, even if the tag <identifier> exists and is mainly related to the DOIs tagging, it is not rare that a DOI is classified as a URL. This happens in particular when the DOI is reported as a searchable link with the "https" prefix. Thus, it is not completely wrong to identify it as a URL. Nonetheless, while in this phase there is no chance to distinguish between these two ways, in the comparison phase this aspect will be kept into consideration.

Table 8, instead, reports the cases in which the analytic and monographic sections are created. In a way similar to the case of Cermine, the co-presence of the <title> and <source> tags allows for the definition of these two sections. The same reasoning applied there are valid also for this section.

Table 7. ExCite metadata and TEI conversion

XML (ExCite)	TEI XML
<author>	<author> <person>
<edition>	-
<editor>	<editor>
<fpage>	<biblScope unit="page"

	from=””>
<given-names>	<forename>
<identifier>	<idno type=”DOI”>
<issue>	<biblScope unit=”issue”>
<lpage>	<biblScope unit=”page” to =””>
<other>	-
<publisher>	<publisher>
<reference>	<biblStruct>
<source>	<title>
<surname>	<surname>
<title>	<title level=”a”>, <title>
<url>	<ref src=””>
<volume>	<biblScope unit=”volume”>
<year>	<year>

Table 8. ExCite analytic and monographic

	+ <title> + <source>	+ <title> - <source>	- <title> + <source>	- <title> - <source>
analytic	X		X	
monograph	X	X	X	X

For what regards Pdfssa4met, things are more complicated. Indeed, it provides an output in generic XML, not guided by specific standards. While the fact that it is written in XML is useful since it allows us to make use of a part of the same structures used for the previously described tools, on the other hand it is completely different in the sequence tagging. Moreover, the software is not able to retrieve all the metadata of the references. The set of potentially identifiable metadata is not explicitly provided in the Pdfssa4met repository, but it can be assumed by the regex list in the code ‘`References.py`’. This list is reported here for a matter of information: *edition, pages, title, URL, volume* and *year*. The authors are not considered, together with the series information, issue and other

potentially relevant metadata. Moreover, from a concrete point of view, Pdfssa4met is able to retrieve only a few metadata from this starting set, i.e., the pages and the date. It is not rare also, that one of the pages is confused with the date. Thus, while the parser is considering all the tagging options, the effectively used ones are just the ones specified here. Probably, since the tool makes use of regular expressions in order to detect the references, a better result could be provided by a formulation improvement. But since that was not the focus of this research, this is more a suggestion for potentially further research.

A first consideration about the conversion procedure concerns the fact that the output is by default printed and not saved in a new file. Indeed, by default Pdfssa4met only prints the output references in the terminal, without creating a new file containing the extracted and parsed references. Nonetheless, a saved output is necessary in order to proceed with the following tasks of this research. Therefore, in order to find a solution, I decided to add to the official code three lines where a new file with the XML citations is created. The following lines have been added between lines 136 and 137 of the original code, where the single references are printed with *sys.stdout.write*:

```
with open('path/to/file/xml', 'a') as file:
      file.write(ref + '\n')
file.close()
```

This is not the only solution, but it seemed the simplest to get an output file.

For what regards, instead, the conversion script file (Cioffi 2022a), the metadata retrieved are scarce. Nonetheless, because of its similarity of tagging with Excite, these two parser outputs were provided with only one conversion script.

Table 9 provides the metadata potentially identifiable by Pdfssa4met. These metadata are reported in the `parsers.py` file, findable in the GitHub repository of the project. All of them are reported with the respective translated node, excepted by the edition element which is one of the metadata never considered as relevant in order to identify the publication types. Thus, it does not require a translation. Also, the URLs are theoretically provided separately from their respective citations, but they are considered because they are actually identified.

Table 10 shows the cases in which the analytic and the monographic sections should be used. In this case there are not enough elements capable of showing the presence of an analytic or monographic section. Also, almost all the metadata retrieved belong to this section (year, pages and title), while data like the authors are not even identified. Therefore, only the monograph element is taken into consideration, excluding the analytic and, of course, the series sections.

Table 9. Pdfssa4met metadata and TEI conversion

XML (Pdfssa4met)	TEI XML
<edition>	–
<pages>	<biblScope unit=”page”>
<reference>	<biblStruct>
<title>	<title>
<url>	<ref src=””>
<volume>	<biblScope unit=”volume”>
<year>	<title level=”a”>

Table 10. Pdfssa4met: monographic

	*
analytic	-
monograph	X

Scholarcy, just like Cermine, is a tool which provides the output in different formats, including Jats. The reason behind the selection of this format is that, while it has all the advantages of the XML language, it follows the general rules stated by the Jats encoding. Indeed, the standard XML, while being able to tag less metadata, is also less formal. At the same time, selecting BibTeX implies the necessity to implement new systems in order to convert that format to XML TEI. Thus, the best output in terms of quantity of information and qualification is Jats. Differently from Cermine, Scholarcy presents a wider variety of tags and a more specific identification of the publication type. Indeed, while in the Cermine Jats output the presence of an article rather than a journal, or of a chapter rather than a book was identifiable through empirical information, i.e., the copresence of the nodes <article-title> and <source>, Scholarcy provides a tag specifically thought to declare the reference type. <element-citation> has this role in this environment. It is accompanied with one of the following tags, declaring the publication type of the cited source: *article-journal, book, chapter, journal, misc, paper-conference, report, webpage*. Thus, in this case it is straightforward to reconstruct the monographic and the analytic sections. In case the attributes article-journal, chapter

or paper-conference are present, both the sections are created; otherwise only the <monogr> one is considered as correct. In a similar way, Scholarcy can distinguish between the author's and the editor's names. This difference is actually relevant for the comparison with the gold standard since it allows to have a specific identity to both those figures. The different meaning is carried, again, by an attribute: <person-group person-group-type="[author | editor]">. Another positive aspect concerning Scholarcy, is that it is able to recognise the <series> section. Indeed, the presence of this tag, absent in all the other references extractors, allows for the creation of a third section, precisely the series one. This section is rarely present and thus recognised by the tool either for a structural (i.e. it is even not considered as an option to be recognised) or concrete (i.e. it is difficult to identify). Here, instead, the tag is reported lots of times and, therefore, it is possible to create the <series> section containing the <title level="s"> node, whose content is the original <series> node. Finally, it is relevant that only Scholarcy is able to recognise "issue", "publisher" and "publication" place as standard metadata. The code of the conversion can be found in (Cioffi 2022a)

Table 11 includes the tags used by Scholarcy in its version 2. This version has been selected since it is more accurate in the metadata tagging inside the single citations (not in the identification of the citations.) than version 1. All the tags are considered, from the structural to the metadata-specific ones. While almost all the metadata have their own correspondent in TEI, <article-title> and <chapter-title> must be flattened in the same figure, <title level="a">. At the same time, <element-citation> is ignored, since there is no correspondence in the TEI document. *Table 12*, instead, shows the cases in which each section should be created in the references. While the monographic section is always present, the analytic section should be inserted only in case the <element-citation> node has one of attribute 'publication-type' specified in the first column. Finally, the book series are not signaled in the <element-citation> but they are, instead, specified in an inner tag inside the reference.

Table 11. Scholarcy metadata and TEI conversion

Jats XML (Scholarcy)	TEI XML
<article-title>	<title level="a">
<chapter-title>	<title level="a">
<edition>	-

<element-citation publication-type="[article-journal | book | chapter | journal | misc | paper-conference | report | webpage]">	-
<issue>	<biblScope unit=”issue”>
<page-range>	<biblScope unit=”page”>
<person-group person-group-type=”[author | editor]”>	[<author> | <editor>]
<pub-id>	<idno type=”DOI”>
<publisher-name>	<publisher>
<publisher-loc>	<pubPlace>
<ref id=”ref_[n]”>	<biblStruct id=”b[n-1]”>
<series>	<series><title level=”s”>
<source>	<monograph><title>
<string-name>	<persName>
<uri>	<ref src=””>
<volume>	<biblScope unit=”volume”>
<year>	<date when=””>

Table 12. Scholarcy analytic, monographic and series

	<element-citation publication-type="[article-journal | chapter | misc | paper-conference]">	<element-citation publication-type="[book | journal | misc | report | webpage]">	<series>
analytic	X		
monograph	X	X	X
series			X

Finally, Science Parse, differently from all the other tools, provides the output only in Json format. Nonetheless, the creation of a new XML TEI file allowed the reuse of some of the structures originally

created for the XML-to-XML conversion files, e.g., the functions for the creation of the new XML, for the time parsing and the creation of the analytic, monographic and series section for each reference extracted. What changes is the approach to the input file and the library required to parse it. Indeed, the "json" library (for Python 3.8) was used with the purpose of parsing the input file. Through it, the file was parsed in order to find the "references" key, whose value was a list of dictionaries containing the references metadata. Then, for each metadata the four available metadata, i.e., a key-value pair, are extracted and converted in XML TEI through the *lxml* library. The only aspect which required particular attention was the 'venue' metadata. Indeed, the venue is a metadata which can be associated with the concept of publisher (Hunter Library Research Guides 2021), either a book, journal or conference/proceedings. The problem is that from the TEI point of view these elements are distributed in:

- `<title>` (the journals or proceedings, which are considered as the publisher themselves).
- `<publisher>` (a book publisher).

These elements are completely different and with a different meaning from the TEI attribution of the metadata perspective. Indeed, while the `<title level="j">` defines the existence of an article and of its respective journal, the `<publisher>` tag does not provide any structural information about whether the work is a book, a book chapter or an article from a proceedings. Thus, it was not that easy to find a way to translate this attribution of the metadata in the gold standard format. The way which seemed the best in order to carry out this task was that of considering the venue as the `<title>` tag, by default. This decision was taken on the basis of the objectively wider presence of the articles and conferences with respect to the books in the entire dataset. In case this is not the correct tagging, i.e. if the cited work is a book and not an article, a further analysis will be made in the comparison phase. Even if this is not the perfect solution to this issue, it was, in this context, the most suitable.

Thus, the presence of a value for the 'venue' key different from 'null', defines the name of the resource containing the cited work. Any other occurrence different from the ones listed below should be considered as wrong. Thus, the rules used in case of the presence of a non-empty 'venue' data, are the following:

- The metadata used to translate the venue is `<title>` and not `<title level="j">` in the monograph section. This is so to avoid possible mismatches between journals and conferences. Indeed, the presence of the title level is not essential in order to measure the similarity between the two metadata in this research.

- The presence of a non-empty ‘venue’ value allows for the creation of an analytic and a monographic section. In the opposite case, only the monographic section is created, considering the absence of the venue as a signal of the presence of a book.

Differently from the venue information, the authors are presented as a list of strings, where each string represents the sequence of the author’s name and surname. In this case they are treated in the same way as for Scholarcy, where the authors data are in the generic tag `<persName>`. Indeed, the scope of the conversion script is not that of identifying the metadata subcomponents in the output file, but rather to convert the created structures to the correct metadata attribution. If in the references parsing task the tool is not able to distinguish the forename(s) from the surname(s), then this aspect should be maintained until the comparison phase. Finally, the ‘year’ and ‘title’ metadata, instead, are pretty straightforward to manage. Indeed, both those keys are used to define exactly what they mean: the year, either the single year or the full date, and the work title, either monographic or analytic. Another consideration about the conversion process is that, while all the other parsers provide an ID to their citations, Science Parse does not provide this information. Thus, in order to attribute an ID to the references in the final TEI XML file, it was necessary to record the position of each entry in the json citations list and attribute it to the id of the references in the output file.

Table 13 records the metadata which Science Parse is able to identify for each reference. For what concerns venue, and as a direct consequence ‘title’, it is reported only the default translation, which is the only one carried out by the conversion script. In case this conversion is not correct the metadata will be managed of consequence in the evaluation phase. *Table 14* represents the cases in which the analytic and monograph sections are created in the converted output file. Differently from the previous cases, it is not possible to consider the absence of the title and the presence of the venue as only monographic since the accuracy of science parse in recognising the title and the venue is higher than in the other cases and since there are cases in which the title is not reported in the reference but only the venue, thus it is important to keep those two fields separate on the basis of their actual value.

Table 13. Science Parse metadata and TEI conversion

Json	TEI XML
“author”	`<author> <person>`
“title”	`<title level=”a”>`
“venue”	`<title>`
“year”	`<year>`

Table 14. Science Parse analytic ad monographic

	+ "title" + "venue"	+ "title" - "venue"	- "title" + "venue"	- "title" - "venue"
analytic	X		X	
monographic	X	X	X	X

4.4 Comparison and Evaluation Script

In this section it is explained how the comparison and evaluation script has been implemented on the basis of the methodology explained in the previous chapter. All the three aspects, i.e., number of correct references, number of correct metadata and number of correct metadata texts, have been implemented on the basis of the selected parameters and on the retrieved parsing tools.

First of all, the identification of the total number of correct references per paper is the third and last step in the paper-paper comparison phase. Only once identified the fields correctness and the number of correct fields, it is possible to identify the number of correct citations. Indeed, the evaluation of this last parameter must be based on the previous results, in particular on the number of correct fields per reference.

One crucial aspect to take into consideration when evaluating the number of correct references identified is: how many correct fields can be considered enough to allow the definition of two references as the same citation? Indeed, it cannot be taken for granted that all the references will have the same structure, i.e., same number of identified metadata with the exact same metadata content. Thus, it becomes necessary to identify some parameters which must be satisfied in order to consider two references as being the same. These parameters, in the current research, coincide with the most frequently observed metadata fields, in conjunction with the presence of other metadata in other bibliographic entries of the citations list. Indeed, one of the most frequent cases in the reference identification, is that the components of a single bibliographic entry are split in two separate references. Thus, in this case it is not possible to consider both the entries correct, but at the same time it would be not totally correct too to consider both as wrong. Indeed, the fact that one citation is not detected as such, must be considered as an error. At the same time, the reference metadata have been identified, thus it would not be totally correct even to ignore them at all. Therefore, the solution

which seemed to settle these extremes, is to define some parameters that must be considered more characterising than the others in the whole metadata set.

The list of metadata required in order to define each reference is the one reported by words below and in Table 15. The required metadata for each entry differ on the basis of the referenced source type. Indeed, they are selected time by time on the basis of two factors, the publication type and the references parsing tools' features. Another variance factor is the possible lack of metadata in the article itself. In this case, for some publications, an alternative is proposed. For instance, the articles can be identified through the article title or, alternatively, through the combination of volume and pages. This attribution of metadata is reported in a small cluster of the input dataset. Differently, for instance the websites which do not present enough metadata to interchange them in case of lack of the preferred metadata, no substitution is provided. If the title, theoretically required in order to be identified, is not reported, only the URL will be considered. Indeed, even alone it can identify the cited resource. The identification of the metadata has been structured in the following way:

Metadata available in all the bibliographic references:

- *year* (correct if it is retrieved in the exact section).

Metadata available based on the specific publication type (always after checking if the specified metadata is present in the gold standard XML file):

- Journal article: *journal, [article | volume, page]* (at least one of the two sets of metadata between squared parenthesis must be considered on the basis of the retrieved metadata and of the present metadata). The article title is the best option to identify the work. But in a few papers the titles are not provided (e.g., in the Royal society of Chemistry, see the file named CHE-ENG_10). Thus, in these cases it is necessary to identify both the volume and the page, otherwise there is no chance to recognise the correct work, since two or more papers may be present in the same journal of the same year.
- (E)book/ report chapter, proceedings, conference: *chapter title, monographic title*. The main metadata set necessary to identify all these kinds of publications is composed by the two chapters, the inner work and the container one (book, proceeding etc.). The remaining information, e.g. author, pages, is not essential to identify it, even if useful.
- (E)book, manual, database, preprint, report, software, standard, preprint: *work title*. Differently from the chapters and the journals, which basically are a work in a work, these

works are independent. Thus, the only elements necessary to identify them are the title and the year.

- Forthcoming articles, unpublished, grey literature: *title, note*. A fundamental aspect connected to this topic is the note. Indeed, without it, it would not be possible to check them perfectly.
- Newspaper: *year, work title, newspaper title.* Just like the journal articles and the book chapters, the newspaper is represented as an article in a journal or magazine. Thus, the elements necessary to define a reference to a newspaper are the two titles: the work's and the magazine's ones.
- Patent: *title, patent number*. Also in this case, the fundamental element to define the identity of a reference to a patent is the patent number. In case two references cite the same patent number they are surely the same citation. In order to guarantee an extremely sure identification, also the patent title is considered.
- Series: *work title, series title.* In the case of the series the series title is required in order to identify the citation. Together with it, the work title is necessarily identified.
- Website: *title, URL*. The fundamental aspect of a website is its URL. Thus, in order to identify a website citation, the URL in connection with the web page title is required. Nonetheless, also the title is a fundamental element to identify a resource on the web.

Based on the specific tool's extraction features:

- Monograph title – not identifiable by Cermine and Pdfssa4met.
- Note – this metadata can be identified only be GROBID and Anystyle. For all the remining tools it should not be considered.
- Patent Number – identifiable only by Anystyle and GROBID. For all the other tools only the remining metadata will be enough.
- Series title – this metadata is not identifiable by Cermine, Pdfssa4met and Science Parse.
- URL – this metadata is not identifiable by Cermine and Science Parse. For these tools only the web page title should be considered.
- Volume and page - they can be found by all the parsers, excepted by Science Parse. In case the journal is not provided there is not an alternative metadata to change with it.

Table 15. Metadata identifiable by each references extraction tool.

	Anystyle	Cermine	ExCite	GROBID	Pdfssa4met	Scholarcy	Science Parse
journal article	*year, journal, [article | volume, page]*	*year, journal, [article | volume, page]*	*year, journal, [article | volume, page]*	*year, journal, [article | volume, page]*	*year, [article | volume, page]*	*year, journal, [article | volume, page]*	*year, journal, article*
(e)book/ report chapter, proceedings, conference	*year, chapter title, monograph title*	*year, chapter title*	*year, chapter title, monograph title*	*year, chapter title, monograph title*	*year, chapter title*	*year, chapter title, monograph title*	*year, chapter title, monograph title*
(e)book, manual, data sheet, database, preprint, report, software, standard, preprint	*year, work title*	*year, work title*	*year, work title*	*year, work title*	*year, work title*	*year, work title*	*year, work title*
forthcoming articles, unpublished, grey literature	*Year, title, note*	*Year, title*	*Year, title*	*Year, title, note*	*Year, title*	*Year, title*	*Year, title*
newspaper	*year, work title, newspaper title*	*year, work title, newspaper title*	*year, work title, newspaper title*	*year, work title, newspaper title*	*year, work title*	*year, work title, newspaper title*	*year, work title*
patent	*Title, year, number*	*Title, year*	*Title, year*	*Title, year, number*	*Title, year*	*Title, year*	*Title, year*
series	*year, work title, series title*	*year, work title*	*year, work title, series title*	*year, work title, series title*	*year, work title*	*year, work title, series title*	*year, work title*
webpage	*Title, URL*	*Title*	*Title, URL*	*Title, URL*	*Title, URL*	*Title, URL*	*Title*

All the above listed parameters are considered as relevant only in case they are found in the gold standard. Otherwise, only the remaining ones will be considered so. The table is indicative for the metadata theoretically necessary to identify, minimally, the reference identity. In case they are not provided by the papers themselves of course they can't be taken into consideration. The remaining ones should nonetheless be enough in order to define the cited works.
Just to make an example, we can see the pages process:

1. Check if the metadata <page to="*page-number*"> exists;
2. Check if the metadata <page to="*page-number*"> exists but it is empty;
3. Check if the metadata <page to="*page-number*"> exists but it presents the same value as <page from="*page-number*">.

Only in case the first answer is true and the second and the third are false we will consider the metadata, otherwise it is ignored in the count of the total metadata of both the files and it won't have consequences in the precision and recall evaluation.

In case all the required metadata are retrieved in the reference extracted by the tool, the two <biblStruct> elements are considered as the same. These parameters clearly allow for non- rigid parameters for the references matching. The reason behind this choice, apart from the ones listed above, lies in the fact that the cases in which not all the metadata are available are so many that if wanting to be too strict in the matching phase only a few references could be matching, even if effectively being the same.
From a practical perspective, in order to cope with this division, each reference in the gold standard dataset has been provided with a specification of its type. In fact, TEI allows for the specification of the type of reference through the key value pair 'type', 'one of the types used in this research (see previous chapter)'. Through this signal, it is possible to identify the publication type of the references source and, thus, select the metadata necessary to make the comparison.

The second level of evaluation is the number of correct metadata for each bibliographic entry. Indeed, one interesting aspect in the evaluation is understanding whether inside one single reference a metadata is wrong because it has not been identified at all, or if the reasons for the missing identification can be attributed to a lack of precision in the metadata identification. For instance, if the correct metadata is <title level="a">This is the title</title> but the parser gets this result <title level="a">This is</title> than there would be no chance to confirm that the two titles are the same title. Thus, from a correctness point of view the value would

be 0. Nonetheless, the metadata ‘title’ with level ‘a’ has been identified. The reason behind this level of analysis is merely analytic. This level allows us to identify the tools which are able to identify the metadata and the ones which are not. The correct creation of a metadata should be considered as a relevant result with respect to the evaluation of the software, but, of course, not from a full correctness perspective, an element which is guaranteed by the metadata content correctness level. In this case, different from the other two analysis levels, the metadata content and the full reference, there are no specific rules to consider. The steps to follow are four, and each of them produces an output that will be used to evaluate the precision and recall of each paper. The steps are the following:

1. Create the three variables that will contain the three values used for the computation of precision and recall: one for the total amount of metadata in the gold standard, one for the total amount of metadata in the output file and one for the total number of correctly identified metadata. Then, enter all the references one by one and:

2. Identify and list all the metadata for each reference in the gold standard. The output is a variable identifying the total number of metadata reported in the gold standard. It is a positive number, computed by attributing the value 1 to each attribute of each reference. Then, the value of each item is summed to the values of the other metadata retrieved. Add this value to the first variable created.

3. Identify and list all the metadata for each reference in the output file. Only the elements included in the range of metadata identified in the gold standard should be taken into consideration. Indeed, potential metadata not present in the gold standard should not be considered since they cannot be evaluated. This aspect is derived from the fact that not all the metadata are considered in order to evaluate the reference quality (see chapter on the metadata selection for the gold standard). Thus, all the metadata that are not considered in the gold standard should be, thus, excluded from the quality evaluation. Vice versa, the items present in the gold standard but absent in the output will be considered wrong. The output of this step is a numeric variable defining the total number of metadata reported in the file to compare. This number can be either greater, equal or smaller to the gold standard one, but always positive or at least equal to zero. Add this value to the second variable created.

4. For each reference, compare the listed items. The difference of this comparison with respect to the other two steps is that there is no margin of error. The element is either present or not. Thus, the evaluation is based on attributing 0 to the missing nodes and 1 to the present nodes.

This last step output is again a numeric variable showing the total number of correctly identified fields. Add this value to the third variable created.

Once these three global variables are identified, it is possible to compute precision and recall for this step for each extraction tool. This will be seen in the next chapter, about the evaluation.

As concerns the verification of the field correctness there are different aspects to point out before explaining the way in which the similarity was measured. First of all, from a semantic perspective, each field has some properties which make it different from almost all the other fields. Indeed, while some fields may be really close (e.g., the starting and the end page of the cited resource), some others are really different from a conceptual point of view (e.g. the title and the personal names). Therefore, in order to make an equal comparison, these specific aspects concerning the single fields should be taken into consideration in order to decide on which rules the comparison should be made on.

For instance, while the title is a single block always identified in the same way, the authors' and editors' forenames are treated differently on the basis of the peculiarities of the different citation styles. Indeed, in some cases there will be all the entire names, while in some other cases only the names' initials are considered. Thus, for instance, it should be decided how to compare the full names and the initials which are not separated by a dot or a comma or a space. But this road could be really hard since it is really difficult to disambiguate between the initials "AL" and the name "Al". In order to avoid these kinds of misunderstandings during the comparison and to simplify the research the comparison process itself only the first initial of the first name (the first letter which follows the surname or the first letter after the previous comma) will be considered as a parameter. Also, for what concerns the generational names like "Jr" or "III", there is a debate about whether they should be considered part of the legal name (for practical reasons) or not (historical reasons)[52]. Nonetheless, actually they are not explicitly mandatorily considered as part of the legal name and therefore I chose not to consider them as a barrier aspect in the output evaluation. In conclusion, a software obtains a high score if it is able to correctly identify the entire full surname and, at least, the initial of the first forename.

A second aspect to consider, from a strictly linguistic point of view, is the necessity to find one or more parameters on the basis of which define the level of similarity between two strings. This necessity derives from the possibility of having different interpretations or encodings of the same string by the parsers. Identifying some predefined and justified parameters to determine whether two

[52] "Junior", *West's Encyclopedia of American Law*

or more strings are the same basically means to retrieve the features that two or more references should present in order to be reasonably considered the same string. The sets of rules defined to make this comparison are two: the preprocessing operations and the similarity approach.

Before identifying a similarity approach to compare the strings, it is necessary to define some formal rules to apply to the texts before comparing them. This operation is required in order to avoid the presence of syntactical errors or non utf-8 characters which may mine the good result of the texts' comparison. Thus, after the identification of the two strings to compare, the one of the parser outputs and the gold standard one, the first step is to trim both the strings. The steps required in order to avoid errors in comparison deriving from non-relevant textual elements are listed in different papers about how to compare two strings. All of these steps refer to both the strings, the gold standard and the output ones.

1. Split the text into single words, in this way the following operation can be carried out on each single word identified.
2. Lowercase the text, in order to avoid problems related to case sensitive processes. Also, since some parsers have access to databases in order to get the final results it is possible that the terms are capitalised if they were not or vice versa (Brownlee 2017).
3. Remove non-ASCII characters. This step is necessary to compare strings avoiding the errors derived from the unrecognized non-ASCII characters.
4. Remove punctuation and over spaces. In this way it is possible to compare the texts avoiding taking into consideration, for instance, words separated by a new line or not correctly removed separator dots, e.g. "bio. gen. mol." != "bio gen mol", (Khalid 2020). Also, this option is carried out in all the cases excepted by the URLs and DOIs for which the syntax is fundamental in order to identify the resources.

These operations are carried out with the NLTK Python library, specifically created with text analysis purpose. The code is available in the source (Cioffi 2022a). Once the strings are trimmed, the actual comparison phase happens. And here the string similarity rules come in.

First of all, for what concerns the string similarity measures, it is necessary to specify that each field requires separate considerations. Indeed, all the fields have different features and requirements. Thus, for each of them it is necessary to verify all, or at least the most probable, cases and exceptions in order to define a metric and the related rules. There are three major ways to measure the string similarity: String-based, Corpus-based and Knowledge-based (Vijaymeena and Kavitha 2016). Nonetheless, in the current case, the most suitable way to obtain the level of similarity between two

strings, out of a corpus and without enough a priori knowledge, is through a string-based similarity. Also called 'string distance function', the string metric is a metric which allows us to compute the distance, or numerical difference, between two strings. Despite the distance being a measure opposite to the similarity, in this case it is the key to verify whether two strings are similar enough to be considered the same or not. Indeed, by setting a maximum distance value, the similarity is measured by verifying whether the output value is smaller or greater than the selected threshold. In the former case the similarity is verified, in the latter the two strings cannot be considered sufficiently similar, and, thus, different. There are three main categories of string metrics functions: edit distance, token distance and hybrid distance functions. The edit distance functions are based on the computation of the minimum number of operations (addition, deletion, replacement or transposition) required to transform one string into another one. There are different measures (Cohen, Ravikumar, and Fienberg 2003), each with different features, e.g. the Levenshtein distance (addition, deletion and replacement), suggested for short strings;[53] the longest common substring problem (addition and deletion), suggested for cases in which the interest is on specific sequences where the characters are located in consecutive positions and the Jaro-Winkler distance (transposition),[54] based on the idea that two strings are more similar (less distant) if they are similar from the beginning characters. The token distance functions start from a different point. Indeed, from the token distance function perspective the strings are considered as "multisets of words". Some examples are the Jaccard similarity, the TF-IDF (term frequency - inverse document frequency) and the Jensen-Shannon. Finally, the hybrid distance functions, as the name suggests, are a mixture of token based and edit based distance functions. An example is the recursive matching scheme like the Monge-Elkan (Cohen, Ravikumar, and Fienberg 2003). This similarity measure is suggested for long strings.

In case these measures provide a result as a specific number (e.g. it is the case of the Levenshtein measure), it must be normalised in order to be evaluated. Normalising the result is necessary in order to compare the level of the difference between two strings independently from their length. The normalisation process, in case it is a separate process from the main distance computation, takes place right after it. Thus, the values are converted in a number between 0 and 1, with 0 meaning mismatch and 1 perfect match.

A second parameter to take into account is the definition of the similarity threshold or delta (δ). The delta is the minimum level of similarity under which two strings cannot be considered as the same. In this study the value of the δ is, again, a normalised measure. Each typology of data has been

[53] https://www.nltk.org/howto/metrics.html
[54] https://www.nltk.org/howto/metrics.html

provided with a specific threshold, based on the level of similarity required by each data type in order to identify another string to be considered as the same. The values have been selected in a range between $\delta = 1$and $\delta = 0.85$ on the basis of the data specific requirements. This parameter is associated with the normalised value of the distance function. For each type of metadata, it has been presented a theoretically different selection of distance metrics and deltas, on the basis of the specific requirements of each of them. Concretely, the Levenshtein distance, for different reasons depending on the data type, has been identified as useful distance metrics for all the metadata, while the deltas are different.

The **URL** is a kind of metadata that should not be excessively cleaned before the comparison. Indeed, the parts composing it are relevant in order to identify it as a link describing the location of a resource. Nonetheless, since it is a type of string composed of multiple subsequences separated with different punctuation elements, depending on the substring, a small error margin is accepted since the chances of not getting it fully are consistent. Indeed, the only acceptable errors are the missing of an initial or ending part of a word. No inner modifications are acceptable. Thus, while theoretically an evaluation like the Longest Common Substring could be useful, in this context the Levenshtein is enough. Indeed, the aim of this comparison is to check whether the tool was able to correctly parse an input string containing a URL and not to check if two different strings which are not related by a common parent string. Thus, the probability of matching two identical strings with an inner difference due to the tool's error is almost impossible. Thus, the Levenshtein is the selected measure to evaluate the entire string with the single errors, rather than just a substring. To make an example, "*https://www.cs.cmu.edu/~wcohen/postscript/ijcai-ws-2003.pdf*" is the same as "*www.cs.cmu.edu /~wcohen/postscript/ijcai-ws-2003.pdf*" or to "https://*www.cs.cmu.edu/~wcohen/ postscript/ijcai-ws-2003*". Instead, whether it is different from the original string because of a missing piece, e.g. "https://*www.cs.cmu.edu/ ~wcohen/*" or for being completely wrong, e.g. "https://*www.cs.cmu.edu/ ~wcohen/ postscript/something_else2005*", or if it is empty, it must be considered as wrong. The Levenshtein distance is the most appropriate similarity measure for this kind of analysis. Indeed, because of the miscellaneous composition of the websites URLs it is able to analyse all the operations which may be required to make a good matching between two strings of this type. The similarity threshold is set to 0.95, high enough to ignore small imprecisions which are acceptable for this kind of data.

Instead, the **dates** and the **DOIs** are two different data types which share the same necessity for a high level of precision in order to be considered correctly identified. The distance between the dates has its focus on the year. This means that, while the day and the month can be considered as relative

aspects, thus they should not be correct or present in order to identify the date, the year must be correct. Indeed, as for other metadata, especially the numerical ones, the perfect matching of the value is necessary to correctly identify the correct reference. Indeed, on the one hand since the operation required is copying and adding a value to data, there is no scenario in which 1995 could be considered as a slightly wrong identification rather than a completely wrong one. On the other hand, a slightly different date could identify another reference. Thus, it is not possible to define a correctness range for this kind of data. Just to see an example, If the correct date is "1997-28-11", only the possible outcomes "1997" and "1997-28-11" can be considered as correct. All the other occurrences, e.g., "1995", "997-28-11", "199", should be defined as wrong. For what concerns the DOIs, this kind of metadata is really specific and requires a high level of precision in order to be identified. Indeed, one single number can make the difference between two different resources. Since this kind of metadata has a precise structure, there are some steps to follow in order to define it as the correct DOI. First of all, it is necessary to check the presence of the structure "10.". If this structure is in the identified DOI, it is used as a splitting point for the two parts of the string, the one that precedes and the one that follows it. The next step is checking if the numeric part which follows the "10." and the part preceding it are the same as the ones in the gold standard. If all these passages give a positive result, then the two DOIs can be considered the same. Instead, if even one of them is negative, then they cannot be considered as the same, and the output must be signed as wrong. For both these cases, again, the Levenshtein distance is the most appropriate similarity measure, but this time the threshold is set to 1. In this way it is possible to avoid even the small imprecisions and two dates or DOIs are the same only if they are perfectly coincident.

The same considerations made for the dates should be applied to the **issue**, to the **volume** and to the **pages**. Indeed, unless it is obtained a good match of the volume, on which basis "Science vol. 23 issue 4" and "Science vol. 24 issue 4" could be reasonably considered as the same journal? The only way for doing this could be checking the rest of the reference to verify whether the other data coincide. But this operation is not possible in all cases. In fact, in some papers the volume and the pages are the only way to recognise a paper, in absence of the article title, check Table 15 for further information. Thus, the required precision for these three metadata is high, because of their relevance in identifying the referenced papers, but nonetheless it is not required a perfect match in all the cases. Indeed, a minimum margin of error is acceptable since the probability of confusion between two or more strings diminishes with the increasing of the numbers composing the volume or page range (simply from a combinational point of view). By stating the delta equal to 0.9, only for the pages, volumes or issues with at least 6 characters it is accepted one character error. Indeed, only the 10% of 6 provides a value higher or equal to 0.5 which, rounded upwards, results 1. To make an example:

if the volume (or the issue or the page) is “123”, only “123” can be considered as correct. Potential deletion like “12” or “23” would point to different volumes (issues or pages) of the same journal. The same can be said for completely wrong numbers like “4” or “125” or the absence of a value, “ ”. Instead, in a case like “105346”, “10534” and “05346” can be accepted. The selected measure is the Levenshtein distance. The reason for this choice relies on the fact that all the three data typologies are based on one single string composed of numbers and in some cases utf-8 based letters. Thus, there is no chance that changes in the inner part of the number could happen. This aspect is fundamental from the perspective of the identification of the date: from a correctness point of view, it is less wrong the string “193a” than “183”, given as gold standard “193”. While the first is a symptom that the date has been wrongly parsed, in the second case this is surely the wrong page. At the same time this solution avoids the inference of other possibly retrieved characters.

The **patent numbers** are codes represented by a single string. For the way in which they are meanly structured the distance method should accept a margin of error. In fact, patents are cited differently on the basis of the citation style but also on the patent type. To make an example, the following three patents are all cited in APA style. “U.S. Patent No. *10,788,482*”, “European patent No. *0673422B1*” and “U.S. Patent Application No. *2002083598(A1)*”. Such a variance in only the patent number requires a quite elastic measure and a not too rigid level of accuracy. To make an example, in a string of 15 characters like “US20200102230A1” it is accepted a number of wrong characters of 2. It can be either deletion, e.g. “US20200102230A”, addition, e.g. “oUS20200102230A1” or substitution, e.g. “US20200102230AI”. For short strings with a discrete variance like the patent number, the Levenshtein distance is a good measure for the correctness level. Indeed, it guarantees all the previously mentioned editing in a straightforward way. The threshold is set to 0.9, a percentage in line with the other parameters with similar requirements.

For what concerns the **unpublished**, **thesis** and **forthcoming notes**, the most relevant aspect is that the concept of ‘unpublished/ thesis/ forthcoming work’ should be reported. This means that a tolerance range should be included in the evaluation. Indeed, there is no standard way of recording an unpublished or forthcoming article or a thesis. The same concept can be reported differently in two different articles on the basis of the citation style or simply on one of the alternatives offered by the same style. It is the case of unpublished and forthcoming works in the APA style. The official guidelines report these formats: “*Unpublished manuscript [or "manuscript submitted for publication," or "Manuscript in preparation"]*”. An alternative found in the dataset papers is the simple “unpublished” note. The same is true for the other occurrences. Example: “Unpublished manuscript” can be associated with “unpublished” or to “unpublished manuscript”. "manuscript submitted for publication", in the same way, can be considered as enough with “submitted for

publication" or simply "submitted". Thus, when comparing two outcomes, it is necessary to verify that at least the keyword root is correctly reported. Then, even if the two phrases per se have a low similarity, in any case not below $\delta = 0.85$, the match should be guaranteed. The keywords are provided in a list. In case one of them is found in the note to compare then, this principle will be applied. In the other cases the Levenshtein is applied with a threshold of $\delta = 0.85$.

A similar reasoning can be made for the titles at all the levels, journal, books, series, articles and chapters. Their features are almost similar, long sequences of words in the same string. One specific attribute of the journals is that the title could be in abridged format. Thus, only the initials of the words composing the title are reported. For instance, with respect to the example title "Knowledge Management in Software Testing" the forms "Knowledge Management in Software Testing", "Knowledge Management in Software", "Management in Software Testing" can be accepted. Instead, "Knowledge Management in Software Testing Engineering Journal", "Management in Software", "The Knowledge in Knowledge Management" and an empty string are considered wrong. In case the title is a journal title, e.g. "e-Informatica Software Engineering Journal", also the string "EISEJ", the abridged form, should be accepted. Again, the distance function selected for this case is the Levenshtein distance with $\delta = 0.85$. In this way it is possible to check whether the tools have done a good job at parsing the titles leaving a margin in particular for possible extra tokens before or after the correct title.

Finally, for what concerns the author, editor and publisher metadata, again the Levenshtein distance has been applied to measure the similarity. The main reason, this time, is related to the fact that the Levenshtein is really useful when comparing short strings, like the personal names are. To make an example, in case the correct name is "Indrawati" the acceptable results should be the perfect match "Indrawati" and reduced misspellings or missing characters, e.g., "Indrawat" and "ndrawati". Instead, completely wrong names, e.g., "Jordan" and "Indraw" and the absence of a value, an empty string, should be considered wrong. The delta in this case is set to 0.85. Also, in this case there should be a margin for elasticity since the names are generically short and allowing for small changes, again as extra characters at the beginning or at the end of the word, where present, means providing a low range.

5. Results

In this section are reported the results obtained through the different steps in which the papers belonging to the original dataset have been modified. The steps in which an actual analysis on and modification of the data has been carried out are three: the results obtained in the extraction and parsing tasks carried out by the selected tools; the results of the application of the conversion-to-TEI scripts to the parsed results; and the results of the final comparison with the gold standard. The first paragraph regards a generic overview of the results obtained through the extraction and parsing phase for each tool. In particular, possible anomalies observable in the output files are investigated. Second, the results of the conversion phase are described. In this case the focus of the analysis is on the effects of the conversion scripts on the original output files. Which means to verify whether there were changes in the structure of the file or if it remained almost the same. Finally, the results of the comparison and evaluation are reported, both on the entire dataset and on the single scientific fields which it is composed of. In particular, the analysis of results reported in this last section will be postponed in the discussion section.

5.1 Results of the Extraction and Parsing Tasks

As regards the results derived by the conversion and parsing phase, the focus is on the possible problems or good results obtained, in particular if these differ from the original expectations. As concerns the problems detected, there are two different identifiable classes of issues: problems which can be attributed to more than one parser on the basis of common features and problems related to single parsers. An instance of the former issue is the impossibility for the online tools, i.e. ExCite and Scholarcy, to parse large sized files. Indeed, in both cases the same paper (weighing almost 37 MB) couldn't be converted and a server error message (error 503) was returned. This is a signal of the fact that, whether willing to parse large files, other solutions should be taken into consideration. This could mean either to download the source locally, in the case of ExCite, or to request for more available space, in the case of Scholarcy. Indeed, on the one hand all the other tools, working locally or with a locally created server, were able to parse all the files. Thus, we can suppose that also for ExCite this problem could be overcome by downloading the software locally. On the other hand, Scholarcy is only available as an API. Consequently, a request for more space, probably behind payment, is the only available option. Another issue belonging to this group of errors is the impossibility to retrieve any reference from one or more PDF files. This is a widespread problem for

which different explanations can be found, on the basis of the way in which the tool works. The first case to point out is when no "References" section is found. Indeed, two articles have been inserted in the dataset in order to provide a view also on this aspect ("z_notes_test_1" and "z_notes_test_2"). The statistics on the results, before actually checking the quality of the parsing, are good. Indeed, out of seven identified tools, only one, Pdfssa4met, did not identify the section in both the articles, and another one, Science Parse, identified the section in only one out of the two papers. All the other tools identified at least one reference for each of these two papers. A second case regards the fact that some of the tools were not able to extract references from certain papers, even if they had the "References" section provided. In some cases, these are random papers non identified by one single parser, e.g. the second paper of the chemical engineer field has not been parsed by Science Parse while the second paper belonging to the health professions has not been parsed by ExCite. In some other cases instead, two or more tools were not able to parse a same file. It is the case of the papers in the chemical field, where one of them has not been parsed, the first of them by Pdfssa4met and the other by Cermine and Science Parse. By the same tools the papers named "DEN_17" from the dentistry field, "MUL_39" from the multidisciplinary field and/or "NUR_44" from the nursery research field[55] have not been recognized. The first thing to notice in this set of not parsed references is that none of them had the references written in one column layout. From this observation it can be assumed that for the papers in which the references are provided with in a one single column layout, there is a higher probability of having them parsed with respect to a two (or three)- columned page layout. Instead, the fact that Science Parse has not parsed both the files with a three-column structure is a sign that it may have some difficulties in working with this kind of files. Of course, two files are not enough to state this with certainty or to unequivocally identify an issue, but they can be a sign of it. A further investigation may provide more details about this aspect of Science Parse. For the remaining articles in the dataset no particular aspect can be identified that may justify such a lack of identification of its inner references, apart from the fact of being written in a two or more columned layout.

As concerns the second type of error, regarding the issues related to the single parsers, there are two main occurrences, mostly related to the occurrence of badly tagged metadata. The first type of error identified regards the metadata identification by Cermine. Indeed, it seems that Cermine is not actually able to identify the structures of the book chapters. Indeed, just like the books are identified as `<title level="a">` the same is true in case both the chapter title and the book title are identified. This is a problem probably related to the capability of Cermine to identify the journals as source and not the books. Thus, this is a problem that will probably be reflected in the evaluation

[55] see Table 2 for the link to the resource.

phase where the book chapters will be recognised but with the book title counted as an error since it is wrongly tagged and there is no way to separate the two titles. As concerns the other tools, errors like this one are not encountered. Differently, Pdfssa4met provides concretely a small part of the potentially identifiable metadata: date, volume and year only. Indeed, in only one parsed paper out of fifty-six the title has been correctly identified. Also, at first sight, it seems that it carries out badly the extraction task for the single references. Indeed, in some files it retrieves only a few references with a lot of years and pages reported in the same references. It is a signal that it considers as one single reference all the references whose years are reported in the single one selected. The sources, instead, are never identified.

Good results, instead, have been obtained specifically by some parsers. With good result in this section, it is made referment to the quality of the parsing at first sight. This is a judgment only based on the quality of parsing before the actual comparison against the gold standard, thus this comment regards the investigation of whether a metadata has been identified or not, rather than if it is correct or not. With this perspective, GROBID and Anystyle had good results since they were able to parse the references section in all the papers, the former providing them in XML TEI and the other in JSON. In the same way, Scholarcy, apart from two non-parsed files because of server errors, identifies the references in all the remaining files. The worse score from the perspective of the number of references section identified in the dataset goes to Science Parse. Indeed, it performed the maximum number of non-parsed papers, which is of six out of fifty-six papers, i.e., the 10,7% of the entire dataset. The remaining tools, Pdfssa4met and ExCite have instead four (7%) and two (3.6%) missing identifications respectively. Only the comparison results will provide more information about the actual extraction and parsing quality of the parsed references.

5.2 Results after the Conversion to TEI XML

This section reports the results related to the conversion phase. This step has followed chronologically the extraction and parsing phases. The results reported in this chapter show the way in which the structured output of the parsers has been modified in order to be compared to the gold standard. This conversion step has not substantially modified the general results obtained in the previous phase and no particularly problematic aspects have been identified during this process. The main modifications carried out on the original output files have been the conversion to XML, the conversion to TEI and the removal of non-pertinent sections of the paper analysed.

The conversion to the XML language in case it was not the original one: this has produced some changes in the structure of the previous files. For instance, Anystyle which provides its output in Json, ordered the authors as a list of dictionaries where the metadata are provided as key-value pairs where both the key and the value are strings. In the conversion to XML both these aspects were missed: the text key "author" which includes the list of dictionaries in which the single authors data are stored, has been converted to an XML tag, <author>, which instead of including all the references in one single structure, defines them as single elements. Then, each dictionary in the list, containing an author's data, is translated with the tag <persName>, inside which the single data (forename, surname and particles) are singularly tagged with the specific element.

```
<author>
  <persName>
    <surname>B</surname>
    <forename>Forkman</forename>
  </persName>
</author>
<author>
  <persName>
    <surname>LJ</surname>
    <forename>Keeling</forename>
  </persName>
</author>
```

Figure 7. Author in xml

```
"author": [
  {
    "family": "B",
    "given": "Forkman"
  },
  {
    "family": "LJ",
    "given": "Keeling"
  }
],
```

Figure 8. Author in JSON

The usage of the TEI conventions has provided a generally more structured definition to the body structure. An example can be viewed in the conversion of the results outputted by Cermine. The main differences applied to the new text are: introduction of the structures analytic-monographic to identify which metadata belong to which item; deletion of non-tagged elements, either punctuation or strings, in order to clean the results; instead of a generic <string-name> tag to identify the authors, the tag <author> has been introduced; in case the dates are followed by a letter, in order to specify which work by the same author in the same year, both the original form and the cleaned one are reported. The only difficulty related to creation of a more structured organization has been found with the software Pdfssa4met. Indeed, in this case only, no way could be found to provide a structure to the outputted references, and only the monographic section was created. Instead, for what concerns the tagging features of Pdfssa4met, in the same way as ExCite, a more formal structure is related to the fact that the references are numbered and identified through an id, while in the output they were

only tagged with <reference> since by default they were returned as single references. For all the other metadata no other problems have been identified.

```
<ref id="ref94">
  <mixed-citation>
    <string-name>
      <surname>WEDIN</surname>
      ,
      <given-names>L.</given-names>
    </string-name>
    (
    <year>1972b</year>
    ).
    <article-title>Evaluation of a three-dimensional model of emotional expression in music</article-title>
    .
    <source>The Psychological Laboratories</source>
    ,
    <volume>54</volume>
    (
    <issue>349</issue>
    ),
    <fpage>1</fpage>
    -
    <lpage>17</lpage>
    .
  </mixed-citation>
</ref>
```

Figure 9. Reference in CERMXML format.

```
▼<biblStruct xml:id="b94">
  ▼<analytic>
    ▼<author>
      ▼<persName>
          <surname>WIGGINS</surname>
          <forename>G. A.</forename>
        </persName>
      </author>
      <title level="a">Music, syntax, and the meaning of ''meaning</title>
    </analytic>
  ▼<monogr>
    ▼<imprint>
        <date when="1998">1998</date>
        <biblScope unit="page" from="18" to="23"/>
      </imprint>
      <title>In Proceedings of the First Symposium on Music and Computers</title>
    </monogr>
  </biblStruct>
```

Figure 10. Reference in TEI XML

Finally, in each file only the references section has been maintained in case some tools by default parse the entire text or other parts of it. It is the case of Science Parse for which the remaining sections, i.e. the metadata of the current paper, the sections and the references metrics, have been removed.

The same is true for the files parsed by Cermine and Scholarcy. In this way the structure is simplified and only focused on the comparison phase.

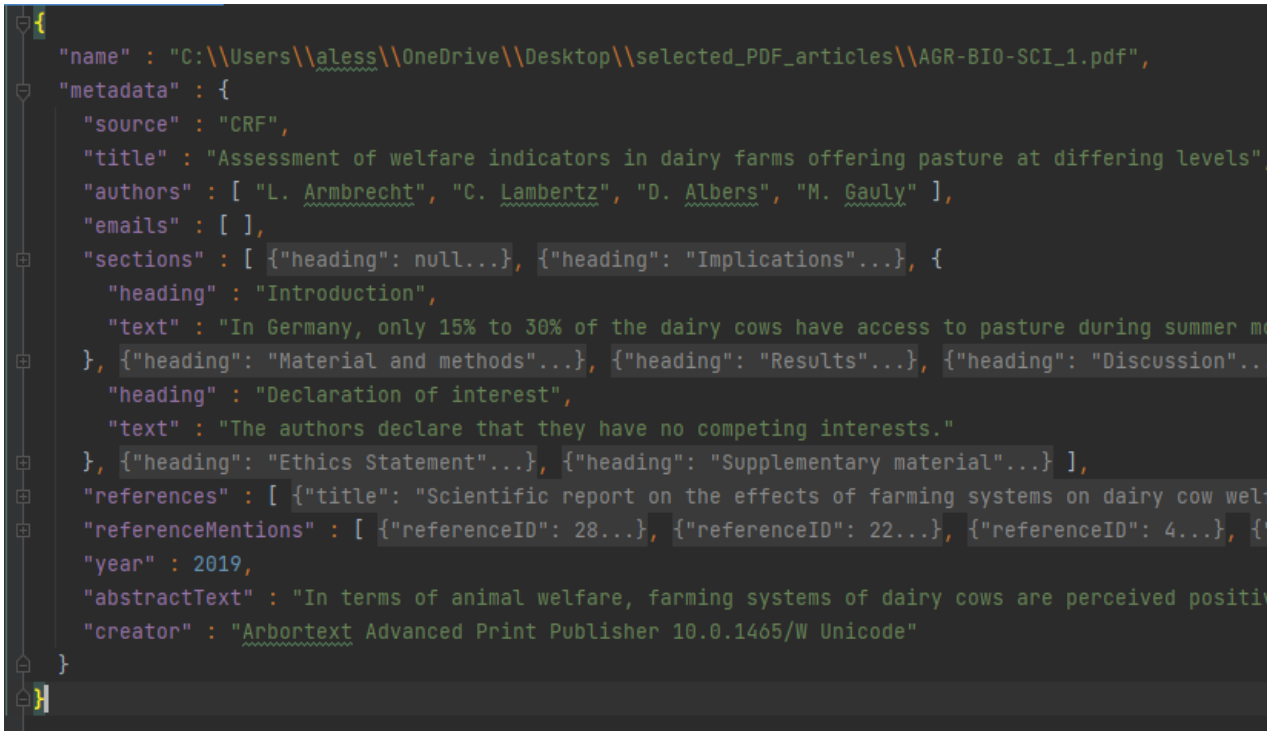

Figure 11. All the sections identified by Science Parse

Figure 12. The only section maintained in XML TEI, the references

Only the GROBID output files didn't require any conversion in order to be accepted, since they are already provided structured in XML TEI. Nonetheless, the lack of this passage generates as consequence the fact that the output files are not cleaned before being compared. This means that the other section of the text, the contents included in the `<body>` element and, inside the `<back>` section, the `<div type="acknowledgment">`, are maintained even in the final file to compare against the gold standard. As consequence, this aspect has to be handled in the evaluation phase, where a specific piece of code has to retrieve the references section excluding the previous text.

5.3 Results of the Comparison against the Gold Standard

In this section the results of the evaluation on each parser are reported. The evaluation has been based on the application of the code described in the methodology and data sections to the output files already converted to XML TEI. In particular, two aspects have been investigated: the results of each parser on the entire dataset and the results of each parser on the single dataset fields. The results of each parser on each single paper in the dataset can be viewed on (Cioffi 2022b). For each of the parsers, the evaluation of the references, metadata and metadata texts levels are reported, rounded to the second decimal.

5.3.1 Overall Results on the Dataset

The quality results of the references parsed by Anystyle show a general high precision level. Nonetheless, it is possible to see a different distribution of the values in the three analysed levels, i.e., references, metadata and contents. The level which received the lowest score is the references one. In particular, the recall on the references shows the lowest score of the entire table. This aspect can be correlated with a low precision of the Anystyle in the identification of enough relevant metadata per reference in order to accept the parsed references and gold standard ones as the same. The precision on the retrieved references is higher, which shows a good intersection between the references retrieved and the total number of correct references retrieved. Instead, the highest level of accuracy is represented by the metadata, whose level is 0.95. This let us understand that the tool is actually good at identifying the metadata composing the reference. Nonetheless, it is less precise in the correct text identification of the metadata. Indeed, the content accuracy level is lower of four decimals with respect to the metadata. As concerns the metadata and the content it is noteworthy the fact that the highest score is obtained in the recall, which means that the number of correctly identified references is closer to the total number of correct references than to the total number of retrieved ones. Nonetheless, Anystyle has a general high f-score in all the three levels.

Table 16. Anystyle values on the entire dataset

	References	Metadata	Content
Precision	0.81	0.93	0.87
Recall	0.74	0.97	0.91
F1	0.77	0.95	0.89

In a similar way, Cermine presents the lowest f-score in the references level and the highest in the metadata level fields. Again, the recall of the references is lower than the precision. Thus, even this tool comes out to be precise with respect to the references identified but it is not good at identifying only the correct references without a certain confusion level. What is noticeable is the attribution of the same (high) value to both precision and recall in the metadata section. In this case these values show that the number of retrieved references is almost the same as the correct metadata. Nonetheless, identifying the same number of occurrences does not guarantee a perfect match, as this value shows. Also, for what concerns the contents, the precision and recall levels are close, but the level is lower which means that the texts are prone to parsing or metadata attribution errors.

Table 17. Cermine results on the entire dataset

	References	Metadata	Content
Precision	0.75	0.94	0.86
Recall	0.67	0.94	0.87
F1	0.71	0.94	0.86

Differently from the previous two tools, the ExCite results present a great difference between the references and the metadata extraction and parsing. The quality of the references extraction is pretty low, even if above the 0.5, and again the ratio of the number of references identified to the number of correct references is lower than the ratio to the number of extracted references. The metadata instead are retrieved with a high level of correctness, with a score a little above the 90% of correct identifications. Nonetheless, the contents inside the metadata are captured with a lower precision. Indeed, on the basis of the same number of identified metadata, the identification of the contents is lower of the 15%.

Table 18. ExCite results on the entire dataset

	References	Metadata	Content
Precision	0.59	0.93	0.79
Recall	0.53	0.92	0.79
F1	0.56	0.92	0.79

GROBID shows a generically good level in the task of the metadata and content elements identification. Indeed, while the references are identified with an accuracy a little higher than the 50%, both the metadata and text are identified with an accuracy higher than 0.85. Also, differently

from the previously described tools, GROBID shows a slightly better score in the recall than in the precision, in all the three levels. Which can be identified with the fact that GROBID identifies more metadata than the necessary ones. Finally, looking at these results it can be said that even if not having a high level of accuracy in the references identification, the contents of the references identified are parsed and labelled with a high precision.

Table 19. Grobid results on the entire dataset

	References	Metadata	Content
Precision	0.54	0.86	0.81
Recall	0.55	0.97	0.92
F1	0.54	0.91	0.86

Pdfssa4met presents the lowest performance among the tools reported. Indeed, its values in the references identification barely reaches the 1% of the entire references in the dataset. Differently, the level of metadata identified outperforms the expectations since it gets to a score of 0.19. In this case precision and recall present strongly diverse values since the recall outperforms the precision of the 50%. This can be associated with the fact that the number of extracted metadata per reference identified is really high, with respect to the total number of correctly identified metadata. Finally, the evaluation of the contents extraction show a general decrease in the parsing quality. Also in this case, the fact that the precision is lower than the recall, confirms the trend already observed in the metadata.

Table 20. Pdfssa4met results on the entire dataset

	References	Metadata	Content
Precision	0.01	0.14	0.07
Recall	0.01	0.29	0.14
F1	0.01	0.19	0.09

Scholarcy presents a trend in the distribution of the precision and recall values which is opposite to the one of almost all the previously presented tools. For what concerns the references, the precision is higher than the recall with an f-score grazing the value of 0.7. The metadata and contents values present a high value in the precision while the recall is relevantly lower. Such a relevant difference of values shows the fact that the number of actually retrieved references is definitely higher than the number of correct references which can be identified. Differently, the metadata and contents show an inverted tendence with respect to the references. Indeed, the number of retrieved metadata is lower with respect to the correct ones, resulting in a low recall. The resulting f-scores nonetheless are high.

Table 21. Scholarcy results on the entire dataset

	References	Metadata	Content
Precision	0.62	0.96	0.90
Recall	0.78	0.70	0.65
F1	0.69	0.81	0.75

Finally, Science Parse presents values in line with the general trend of the parsing values for the single tasks. The references identification value has the lowest F1 score among the three levels, followed by the contents and, finally, by the metadata. A noteworthy aspect is the fact that for what concerns the metadata of the identified references there is a perfect match for all of them, for a precision level of 1. Instead, the recall value is close to the half of the precision one. This clearly identifies a scenario where all the metadata identified are correct, but these are only a few sets of the correct metadata. The same trend can be identified for the contents' correctness analysis value, even if with a generically lower score and a slightly lower distance between precision and recall.

Table 22. Science Parse results on the entire dataset.

	References	Metadata	Content
Precision	0.43	1.0	0.94
Recall	0.32	0.55	0.51
F1	0.37	0.71	0.66

5.3.2 Results per Field

Analyzing the single fields allows to verify at a deeper level the results obtained by each references parser. Of course, since only two files per field have been selected, the results may be influenced by different factors. Nonetheless, this allows to have a further insight in the quality of the parsing where the fields influence the journals. In the following tables the fields are reported in their short form with which they were used in the data analysis. See *Table 2* to check the correspondence between the entire name of the field and its abridged form, used for convenience. In the following paragraphs the results obtained are presented in tables, where the fields in which the top scores have been registered are highlighted.[56]

[56] In the tables from this chapter on, the decimals are represented comma separated. This is so since the programme used to create them used the Italian numeration. However, they should be considered as decimals and not hundreds.

Anystyle shows results which are almost on the same line. All the values for all the fields are above the 0.5 apart from economic finance, where the lowest value in the references reached the value of 0.23. The best results have been obtained in the pharmacological toxic pharmacy for the contentment and the metadata. The best score for the references, instead has been recorded for the business management accounting, with a value of 0.97. Another noticeable aspect is the high quality of the identification of the set of files called 'Z-NOTES-TESTS', the ones which did not present an explicitly named references section. Indeed, all the 'Z-NOTES-TESTS' values lie above the 0.85, results which outperforms other fields in which higher values could be expected before the evaluation phase, e.g. COM-SCI or NUR.

Table 23. Anystyle results for references, metadata and content

	ANYSTYLE								
	references			metadata			content		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
AGR-BIO-SCI	0,92	0,77	0,84	0,97	0,87	0,92	0,84	0,76	0,80
ART-HUM	0,95	0,94	0,94	0,97	0,97	0,97	0,94	0,93	0,93
BIO-GEN-MOL	0,71	0,59	0,64	0,92	0,98	0,95	0,86	0,92	0,89
BUS-MAN-ACC	0,97	0,97	0,97	0,93	0,99	0,96	0,90	0,97	0,93
CHE-ENG	0,87	0,55	0,67	0,82	1,00	0,90	0,80	0,97	0,88
CHEM	0,78	0,73	0,75	0,97	0,99	0,98	0,96	0,98	0,97
COM-SCI	0,80	0,64	0,71	0,78	0,98	0,87	0,74	0,94	0,83
DEC-SCI	0,84	0,85	0,84	0,95	0,99	0,97	0,91	0,96	0,93
DEN	0,90	0,92	0,91	0,97	1,00	0,98	0,94	0,96	0,95
EAR-PLA-SCI	0,86	0,72	0,78	0,90	0,93	0,91	0,48	0,49	0,48
ECO-ECO-FIN	0,93	0,93	0,93	0,92	1,00	0,96	0,89	0,96	0,92
ENE	0,62	0,23	0,34	0,49	0,98	0,65	0,49	0,99	0,66
ENG	0,74	0,55	0,63	0,83	1,00	0,91	0,81	0,97	0,88
ENV-SCI	0,66	0,70	0,68	0,93	0,98	0,95	0,91	0,96	0,93
HEA-PRO	0,89	0,68	0,77	0,89	0,97	0,93	0,87	0,96	0,91
IMM-MIC	0,53	0,59	0,56	0,99	0,99	0,99	0,95	0,94	0,94
MATH	0,74	0,74	0,74	0,96	1,00	0,98	0,95	0,99	0,97
MAT-SCI	0,45	0,40	0,42	0,93	0,91	0,92	0,84	0,83	0,83
MED	0,67	0,49	0,57	0,89	0,94	0,91	0,81	0,85	0,83
MUL	0,73	0,64	0,68	0,97	1,00	0,98	0,95	0,98	0,96
NEU	0,75	0,65	0,70	0,91	0,96	0,93	0,89	0,94	0,91
NUR	0,89	0,80	0,84	0,89	1,00	0,94	0,80	0,90	0,85
PHA-TOX-PHA	0,82	0,84	0,83	1,00	1,00	1,00	0,99	0,99	0,99
PHY-AST	0,80	0,82	0,81	0,96	0,98	0,97	0,95	0,97	0,96
PSY	0,88	0,90	0,89	0,97	0,93	0,95	0,90	0,87	0,88
SOC-SCI	0,85	0,90	0,87	0,95	1,00	0,97	0,93	0,98	0,95
VET	0,95	0,77	0,85	0,93	0,92	0,92	0,88	0,87	0,87
Z-NOTES-TESTS	0,82	0,89	0,85	0,93	0,98	0,95	0,91	0,96	0,93

Cermine, instead, shows in the VET field the best performance for what concerns the references, followed by MED and DEN. As regards the metadata level, the highest f-score is obtained by the MUL field followed by IMM-MIC and VET, while the highest precision and recall values are observed in many different fields. Finally, the references highest scores are recorded for the HEA-PRO, followed, again, by the IMM-MIC one. In this case there is not one specific field overcoming the others but rather, there are different fields with similar values that slightly outperforms the others in one or more among the analysed levels.

Table 24. Cermine results for references, metadata and content

	CERMINE								
	references			metadata			content		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
AGR-BIO-SCI	0,71	0,76	0,73	0,98	0,96	0,97	0,92	0,91	0,91
ART-HUM	0,83	0,81	0,82	0,98	0,95	0,96	0,88	0,85	0,86
BIO-GEN-MOL	0,86	0,88	0,87	0,96	0,94	0,95	0,90	0,88	0,89
BUS-MAN-ACC	0,82	0,80	0,81	0,88	0,94	0,91	0,86	0,93	0,89
CHE-ENG	0,87	0,87	0,87	0,98	0,88	0,93	0,90	0,81	0,85
CHEM	0,78	0,33	0,46	1,00	0,94	0,97	0,90	0,85	0,87
COM-SCI	0,66	0,66	0,66	0,91	0,94	0,92	0,88	0,91	0,89
DEC-SCI	0,73	0,73	0,73	0,97	0,95	0,96	0,91	0,90	0,90
DEN	0,91	0,91	0,91	0,96	0,98	0,97	0,92	0,94	0,93
EAR-PLA-SCI	0,55	0,55	0,55	0,99	0,90	0,94	0,62	0,56	0,59
ECO-ECO-FIN	0,69	0,45	0,54	0,91	0,89	0,90	0,80	0,78	0,79
ENE	0,63	0,63	0,63	0,85	0,96	0,90	0,78	0,89	0,83
ENG	0,53	0,55	0,54	0,95	0,92	0,93	0,88	0,85	0,86
ENV-SCI	0,68	0,68	0,68	0,93	0,96	0,94	0,92	0,95	0,93
HEA-PRO	0,89	0,53	0,66	0,93	1,00	0,96	0,92	0,99	0,95
IMM-MIC	0,90	0,66	0,76	1,00	0,97	0,98	0,96	0,93	0,94
MATH	0,79	0,79	0,79	0,94	0,98	0,96	0,93	0,96	0,94
MAT-SCI	0,68	0,69	0,68	0,95	0,95	0,95	0,89	0,89	0,89
MED	0,92	0,92	0,92	0,80	0,98	0,88	0,78	0,96	0,86
MUL	0,84	0,13	0,23	0,99	0,99	0,99	0,95	0,94	0,94
NEU	0,84	0,84	0,84	0,89	0,97	0,93	0,87	0,95	0,91
NUR	0,78	0,78	0,78	0,91	0,92	0,91	0,89	0,90	0,89
PHA-TOX-PHA	0,83	0,78	0,80	1,00	0,93	0,96	0,91	0,85	0,88
PHY-AST	0,54	0,24	0,33	0,87	0,91	0,89	0,68	0,72	0,70
PSY	0,77	0,75	0,76	0,87	0,92	0,89	0,73	0,78	0,75
SOC-SCI	0,54	0,40	0,46	0,92	0,91	0,91	0,89	0,88	0,88
VET	0,96	0,96	0,96	0,99	0,97	0,98	0,89	0,88	0,88
Z-NOTES-TESTS	0,65	0,72	0,68	0,95	0,95	0,95	0,86	0,86	0,86

ExCite, instead, does not show specific trends in the distribution of the scores. In two out of the three levels under investigation, the maximum level has been obtained by different fields. In the first level, the references one, instead, the BUS-MAN-ACC field has obtained the highest score for precision, recall and f-score, definitely outperforming the remaining fields. Differently, the multidisciplinary field only obtained the highest precision, but the low recall downgrades the f-score. The metadata obtained the highest score in the field of agriculture and biological sciences and neurosciences. The contents reach the top value in the dentistry field followed by the agriculture biological sciences. The lowest level in all the fields has been obtained by the veterinary field where, for different reasons, no reference could be extracted in none of the papers belonging to that field.

Table 25. ExCite results for references, metadata and content

	EXCITE								
	references			metadata			content		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
AGR-BIO-SCI	0,77	0,81	0,79	0,98	0,99	0,98	0,96	0,97	0,96
ART-HUM	0,68	0,68	0,68	0,96	0,97	0,96	0,94	0,95	0,94
BIO-GEN-MOL	0,43	0,42	0,42	0,93	0,96	0,94	0,33	0,34	0,33
BUS-MAN-ACC	0,91	0,91	0,91	0,87	0,96	0,91	0,86	0,95	0,90
CHE-ENG	0,64	0,53	0,58	0,93	0,85	0,89	0,49	0,45	0,47
CHEM	0,43	0,09	0,15	0,76	0,89	0,82	0,15	0,18	0,16
COM-SCI	0,69	0,67	0,68	0,85	0,98	0,91	0,68	0,78	0,73
DEC-SCI	0,73	0,74	0,73	0,96	0,98	0,97	0,64	0,65	0,64
DEN	0,39	0,39	0,39	0,95	0,99	0,97	0,95	0,99	0,97
EAR-PLA-SCI	0,47	0,50	0,48	0,97	0,98	0,97	0,88	0,89	0,88
ECO-ECO-FIN	0,75	0,66	0,70	0,82	0,96	0,88	0,77	0,89	0,83
ENE	0,15	0,17	0,16	0,85	0,97	0,91	0,70	0,80	0,75
ENG	0,55	0,43	0,48	0,86	0,97	0,91	0,52	0,58	0,55
ENV-SCI	0,80	0,73	0,76	0,92	0,95	0,93	0,89	0,93	0,91
HEA-PRO	0,39	0,38	0,38	0,96	0,99	0,97	0,93	0,97	0,95
IMM-MIC	0,51	0,52	0,51	0,99	0,87	0,93	0,96	0,84	0,90
MATH	0,47	0,50	0,48	0,92	0,94	0,93	0,49	0,50	0,49
MAT-SCI	0,49	0,45	0,47	0,96	0,93	0,94	0,65	0,63	0,64
MED	0,06	0,06	0,06	0,91	0,86	0,88	0,83	0,78	0,80
MUL	0,91	0,25	0,39	0,90	1,00	0,95	0,88	0,98	0,93
NEU	0,77	0,73	0,75	0,97	0,99	0,98	0,95	0,96	0,95
NUR	0,83	0,80	0,81	0,88	0,83	0,85	0,85	0,81	0,83
PHA-TOX-PHA	0,62	0,61	0,61	0,99	0,62	0,76	0,63	0,39	0,48
PHY-AST	0,14	0,09	0,11	0,87	0,91	0,89	0,33	0,35	0,34
PSY	0,68	0,76	0,72	0,94	0,98	0,96	0,91	0,95	0,93
SOC-SCI	0,81	0,40	0,54	0,85	0,96	0,90	0,80	0,91	0,85
VET	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Z-NOTES-TESTS	0,61	0,66	0,63	0,98	0,91	0,94	0,77	0,71	0,74

GROBID does not provide a clear trend too. Indeed, the best score for the references is obtained in the business management accounting, but then for the remaining fields its value falls to a lower value. Instead, the highest score for the metadata is registered in a field, immunology microbiology, where the score of the references identified was really low, nearly the half of the best recorded. Nonetheless, its value remains pretty high in the contents level. In this last field the highest score is obtained by pharmacology toxicology pharmaceutics, which in the references obtained a lower score and pretty high in the metadata identification. The f-score values obtained in all the fields, while being different in the number of correctly identified references (0.28-0.85) arrive to a close range in the contents (0.71-0.93)

Table 26. Grobid results for references, metadata and content

	GROBID								
	references			metadata			content		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
AGR-BIO-SCI	0,69	0,64	0,66	0,87	0,99	0,93	0,80	0,91	0,85
ART-HUM	0,62	0,62	0,62	0,82	0,97	0,89	0,78	0,93	0,85
BIO-GEN-MOL	0,58	0,58	0,58	0,88	0,98	0,93	0,84	0,93	0,88
BUS-MAN-ACC	0,86	0,85	0,85	0,82	0,97	0,89	0,79	0,94	0,86
CHE-ENG	0,54	0,53	0,53	0,87	0,97	0,92	0,78	0,87	0,82
CHEM	0,62	0,62	0,62	0,88	0,96	0,92	0,83	0,91	0,87
COM-SCI	0,48	0,44	0,46	0,83	0,98	0,90	0,83	0,98	0,90
DEC-SCI	0,32	0,31	0,31	0,91	0,96	0,93	0,81	0,86	0,83
DEN	0,59	0,58	0,58	0,86	0,99	0,92	0,83	0,96	0,89
EAR-PLA-SCI	0,39	0,39	0,39	0,82	0,99	0,90	0,81	0,97	0,88
ECO-ECO-FIN	0,54	0,47	0,50	0,86	0,95	0,90	0,81	0,90	0,85
ENE	0,29	0,32	0,30	0,89	0,99	0,94	0,82	0,91	0,86
ENG	0,58	0,52	0,55	0,85	0,95	0,90	0,80	0,90	0,85
ENV-SCI	0,59	0,60	0,59	0,86	0,96	0,91	0,83	0,92	0,87
HEA-PRO	0,64	0,64	0,64	0,88	0,96	0,92	0,81	0,88	0,84
IMM-MIC	0,45	0,47	0,46	0,93	0,99	0,96	0,84	0,89	0,86
MATH	0,30	0,26	0,28	0,75	0,97	0,85	0,70	0,90	0,79
MAT-SCI	0,39	0,38	0,38	0,89	0,96	0,92	0,83	0,89	0,86
MED	0,48	0,60	0,53	0,84	0,92	0,88	0,78	0,85	0,81
MUL	0,48	0,49	0,48	0,81	0,97	0,88	0,79	0,95	0,86
NEU	0,47	0,45	0,46	0,93	0,96	0,94	0,70	0,72	0,71
NUR	0,61	0,59	0,60	0,83	0,98	0,90	0,81	0,96	0,88
PHA-TOX-PHA	0,56	0,65	0,60	0,90	0,98	0,94	0,89	0,97	0,93
PHY-AST	0,43	0,43	0,43	0,82	0,99	0,90	0,76	0,92	0,83
PSY	0,83	0,83	0,83	0,85	0,97	0,91	0,80	0,91	0,85
SOC-SCI	0,51	0,52	0,51	0,87	0,94	0,90	0,82	0,89	0,85
VET	0,59	0,59	0,59	0,88	0,98	0,93	0,79	0,87	0,83
Z-NOTES-TEST	0,62	0,68	0,65	0,86	0,99	0,92	0,82	0,95	0,88

As concerns **Pdfssa4met**, the values different from 0 have been underlined. Indeed, only in seven out of twenty-eight fields a minimum of references and metadata has been identified. The range of the precision with which the references have been identified goes between 0.01 and 0.03. The values obtained in the analysis of the metadata increase up to 0.26 but he minimum remains under the 0.1 for the health professions. Finally, while the business management accounting maintains a high level of precision also in the references identification most of the remaining values loose some decimals with respect to the metadata identification values.

Table 27. Pdfssa4met results for references, metadata and content

	PDFSSA4MET								
	references			metadata			content		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
AGR-BIO-SCI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ART-HUM	0,02	0,01	0,01	0,15	0,18	0,16	0,10	0,12	0,11
BIO-GEN-MOL	0,01	0,01	0,01	1,00	0,09	0,17	1,00	0,09	0,17
BUS-MAN-ACC	0,04	0,03	0,03	0,09	0,33	0,14	0,06	0,22	0,09
CHE-ENG	0,03	0,01	0,01	1,00	0,15	0,26	1,00	0,15	0,26
CHEM	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
COM-SCI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
DEC-SCI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
DEN	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
EAR-PLA-SCI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ECO-ECO-FIN	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ENE	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ENG	0,02	0,05	0,03	0,17	0,33	0,22	0,17	0,33	0,22
ENV-SCI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
HEA-PRO	0,02	0,02	0,02	0,06	0,14	0,08	0,06	0,14	0,08
IMM-MIC	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
MATH	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
MAT-SCI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
MED	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
MUL	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
NEU	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
NUR	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PHA-TOX-PHA	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PHY-AST	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
PSY	0,03	0,01	0,01	0,12	0,27	0,17	0,08	0,18	0,11
SOC-SCI	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
VET	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Z-NOTES-TESTS	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00

Scholarcy obtains the best score in the references identification in the VET field, followed by BUS-MAN-ACC and IMM-MIC. Instead, the f-score values of all the remaining fields remain on a close range of values as concerns the references. Indeed, apart from two values, fixed around the minimum score registered (0.39), all the other values are above the 0.58. A similar trend can be observed at the metadata level where the top score is reached by the ART-HUM field and the multidisciplinary fields with an f-score of 0.87 but the minimum value registered is 0.75 in the VET field. Finally, the values recorded for the contents are close to the metadata ones even if slightly lower. Indeed, the values range for the f-score is 0.62-0.86 where the higher values are obtained by tool which had a high score also in the metadata level.

Table 28. Scholarcy results for references, metadata and content

	SCHOLARCY								
	references			metadata			content		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
AGR-BIO-SCI	0,49	0,76	0,60	0,99	0,70	0,82	0,93	0,65	0,77
ART-HUM	0,63	0,91	0,74	0,98	0,78	0,87	0,91	0,73	0,81
BIO-GEN-MOL	0,69	0,75	0,72	1,00	0,65	0,79	0,93	0,61	0,74
BUS-MAN-ACC	0,83	0,97	0,89	0,91	0,74	0,82	0,89	0,72	0,80
CHE-ENG	0,77	0,84	0,80	0,98	0,66	0,79	0,91	0,62	0,74
CHEM	0,69	0,80	0,74	1,00	0,64	0,78	0,98	0,62	0,76
COM-SCI	0,64	0,70	0,67	0,86	0,76	0,81	0,80	0,71	0,75
DEC-SCI	0,57	0,71	0,63	0,92	0,71	0,80	0,84	0,64	0,73
DEN	0,79	0,95	0,86	0,98	0,70	0,82	0,95	0,67	0,79
EAR-PLA-SCI	0,61	0,80	0,69	1,00	0,70	0,82	0,75	0,53	0,62
ECO-ECO-FIN	0,51	0,75	0,61	0,9	0,75	0,82	0,86	0,73	0,79
ENE	0,61	0,57	0,59	0,92	0,69	0,79	0,82	0,62	0,71
ENG	0,55	0,62	0,58	0,94	0,75	0,83	0,83	0,66	0,74
ENV-SCI	0,61	0,80	0,69	0,93	0,73	0,82	0,91	0,72	0,80
HEA-PRO	0,66	0,62	0,64	0,92	0,72	0,81	0,88	0,69	0,77
IMM-MIC	0,83	0,94	0,88	1,00	0,63	0,77	0,95	0,60	0,74
MATH	0,27	0,74	0,4	0,95	0,74	0,83	0,93	0,72	0,81
MAT-SCI	0,32	0,49	0,39	0,98	0,68	0,80	0,90	0,62	0,73
MED	0,68	0,81	0,74	0,89	0,71	0,79	0,84	0,66	0,74
MUL	0,69	0,84	0,76	0,99	0,77	0,87	0,98	0,76	0,86
NEU	0,63	0,89	0,74	0,98	0,69	0,81	0,94	0,66	0,78
NUR	0,74	0,93	0,82	0,96	0,76	0,85	0,93	0,74	0,82
PHA-TOX-PHA	0,61	0,68	0,64	1,00	0,63	0,77	0,95	0,60	0,74
PHY-AST	0,72	0,81	0,76	0,95	0,71	0,81	0,94	0,70	0,8
PSY	0,63	0,96	0,76	0,99	0,72	0,83	0,94	0,68	0,79
SOC-SCI	0,42	0,42	0,42	0,92	0,80	0,86	0,88	0,76	0,82
VET	0,84	0,99	0,91	0,92	0,64	0,75	0,89	0,62	0,73
Z-NOTES-TESTS	0,60	0,82	0,69	0,97	0,75	0,85	0,90	0,70	0,79

Finally, **Science Parse** presents some peculiarities. Indeed, the maximum f-score obtained in the task of references identification reaches the 0.78 in the business management accounting, followed by the humanities and the health professions. It is noticeable the fact that the precision is fixed to 1.0 in all the fields. This is not so unexpected since two factors are cooperating:

- Science Parse can identify only four metadata per reference, author, title, source and year. Thus, the chances to mismatch different tags are drastically reduced.
- Only a few references are identified. In case a reference is identified it is already clear that at least the metadata necessary to accept it (mainly date, article title and journal title) are already correct and only the authors are missing.
- Science Parse does not consider more results for the same data, excepted by the authors, so there is no chance that more metadata are tagged with same element.

Also, behind these factors, there is also the fact that the medium values in the metadata identification are per se high even in tools where the dynamics are not present. The minimum values lay above the 0.75 with the only exception of Pdfssa4met, where nonetheless the metadata arrive to values which outperforms the references ones of the 160% or more. Because of all these concomitating factors, the analysis of the metadata on the Science Parse results does not provide significative insights into the tool qualities. What is relevant, instead, is the analysis on the contents shows a different scenario where the contents of the metadata are not as precise as we could expect. Indeed, always keeping in mind the aforementioned factors, the scores get to low values which reach the 0.52 in the dentistry field.

Table 29. Science Parse results for references, metadata and content

	SCIENCE PARSE								
	references			metadata			content		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
AGR-BIO-SCI	0,37	0,33	0,35	1,00	0,53	0,69	0,93	0,50	0,65
ART-HUM	0,83	0,63	0,72	1,00	0,55	0,71	0,95	0,52	0,67
BIO-GEN-MOL	0,22	0,22	0,22	1,00	0,54	0,70	0,97	0,52	0,68
BUS-MAN-ACC	0,82	0,75	0,78	1,00	0,53	0,69	0,97	0,51	0,67
CHE-ENG	0,59	0,24	0,34	1,00	0,57	0,73	0,88	0,50	0,64
CHEM	0,33	0,08	0,13	1,00	0,55	0,71	0,94	0,52	0,67
COM-SCI	0,22	0,19	0,20	1,00	0,80	0,89	0,98	0,78	0,87
DEC-SCI	0,26	0,22	0,24	1,00	0,59	0,74	0,85	0,51	0,64
DEN	0,62	0,38	0,47	1,00	0,55	0,71	0,99	0,55	0,71
EAR-PLA-SCI	0,31	0,18	0,23	1,00	0,55	0,71	0,73	0,40	0,52
ECO-ECO-FIN	0,49	0,25	0,33	1,00	0,60	0,75	0,92	0,55	0,69
ENE	0,17	0,14	0,15	1,00	0,56	0,72	0,92	0,51	0,66
ENG	0,52	0,33	0,40	1,00	0,57	0,73	0,99	0,56	0,72
ENV-SCI	0,61	0,39	0,48	1,00	0,53	0,69	0,98	0,52	0,68
HEA-PRO	0,73	0,68	0,70	1,00	0,54	0,70	0,96	0,52	0,67
IMM-MIC	0,62	0,60	0,61	1,00	0,53	0,69	0,98	0,52	0,68
MATH	0,18	0,18	0,18	1,00	0,58	0,73	0,87	0,50	0,64
MAT-SCI	0,15	0,11	0,13	1,00	0,54	0,70	0,91	0,49	0,64
MED	0,44	0,43	0,43	1,00	0,54	0,70	0,99	0,54	0,70
MUL	0,48	0,23	0,31	1,00	0,60	0,75	0,94	0,56	0,70
NEU	0,37	0,30	0,33	1,00	0,54	0,70	0,95	0,51	0,66
NUR	0,42	0,20	0,27	1,00	0,54	0,70	0,92	0,50	0,65
PHA-TOX-PHA	0,38	0,37	0,37	1,00	0,53	0,69	0,92	0,49	0,64
PHY-AST	0,14	0,12	0,13	1,00	0,55	0,71	0,97	0,54	0,69
PSY	0,67	0,50	0,57	1,00	0,55	0,71	0,97	0,53	0,69
SOC-SCI	0,08	0,07	0,07	1,00	0,54	0,70	1,00	0,54	0,70
VET	0,24	0,21	0,22	1,00	0,53	0,69	0,87	0,46	0,60
Z-NOTES-TESTS	0,50	0,09	0,15	1,00	0,56	0,72	0,81	0,46	0,59

6. Discussion

The outcomes of the comparison between the output references and the gold standard ones show a complex scenario in which a tool, Anystyle, results to outperform the others. Indeed, Anystyle obtains the best score in all the three levels of analysis selected in this research, i.e. references, metadata and contents. Nonetheless a deeper analysis of the results, carried out from the perspective of the single research fields and the paper layout, has provided some insights into the results of each specific task. In this regard, a few specifications need to be made before having a look at those alternative results. First of all, in case no reference could be extracted from the input PDF file and in case one o more files have not been parsed at all by the tool, these papers have been considered in the count of the extraction task. In this case, only the references of the gold standard have been counted while all the remaining values are set to 0. Indeed, since no reference could be identified in the file returned by the tool, in a cascade mode, no correct references could be identified and, consequently, no metadata was identifiable. This decision has a consequent negative effect on the references extraction task insofar as the lack of retrieved references provides 0 as f-score for the entire file. The tools directly affected by this resolution are: Cermine, ExCite, Scholarcy and Science Parse.

Moreover, another premise has to be made on the procedure followed for the analysis at the references metadata level. As mentioned in chapter *3. Methodology*, certain tools are not able to extract some of the metadata identified as necessary for certain publication types. For instance, Cermine, ExCite, Pdfssa4met and Science Parse are not able to identify and tag the notes, for instance the annotation in which a publication is identified as a thesis or an unpublished. For these tools, the procedure followed in the comparison script is the following: when comparing the selected gold standard reference and the output one, the note (but more generically the non-identifiable metadata) is not considered in order to compare the references. In this way the references can be compared only against the metadata which can be effectively retrieved. In a second moment when counting and comparing the metadata, all the correct ones are considered, included those which the tool is not able to parse. The aim of this procedure is to allow the references to be identified independently from the metadata they are able to identify, and on the other hand, to investigate the quality of the extraction on all the available metadata in order to have a comparison among tools at the same level. Therefore, the tools which do not include in their domain some metadata are disadvantaged in the proportion in which each of the metadata is frequent in the references. Nonetheless, this mismatch is mediumly not too relevant since most of the references is directed to journal articles and books which, in most of the cases represent the tools first target.

6.1 Overall Results

For what concerns the results of the references extraction and identification there are no surprises. *Table 30* compares the results obtained by each of the seven tools selected based on the results obtained in the references extraction task. The values reported can be located in four different clusters, as visible in *Figure 13* where the f-score are reported in a comparative way. The values can be divided into values around 0.7; values between 0.5 and 0.6; values between 0.3 and 0.4; and finally values between 0 and 0.1.

	Precision	Recall	F1
Anystyle	0,81	0,74	0,77
Cermine	0,75	0,67	0,71
ExCite	0,59	0,53	0,56
Grobid	0,54	0,55	0,54
Pdfssa4met	0,01	0,01	0,01
Scholarcy	0,62	0,78	0,69
Science Parse	0,43	0,32	0,37

Table 30. Results of the references for the different tools.

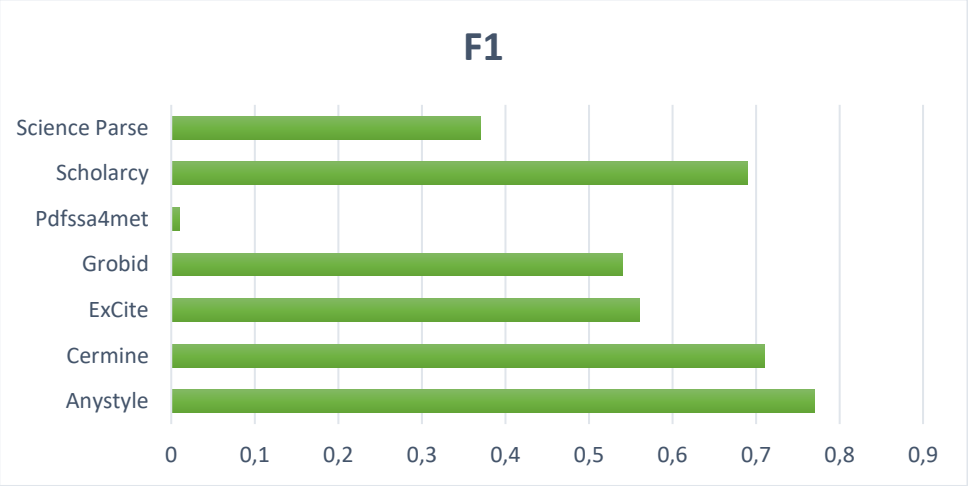

Figure 13. Barchart with the f-score obatined in the references extraction tasks

First of all, the range built around the values of 0.70 includes Anystyle, Cermine and ExCite. The highest values are kept by Anystyle and Cermine, whose values lie above the 0.7. In both the precision and recall values, Anystyle outperforms Cermine of less than ten decimals. Instead, Scholarcy presents the lowest score among the tools in this range, nonetheless close to the Cermine one. In this regard, it is noteworthy the fact that the ratio between the values of precision and recall are inverted

for Scholarcy and Cermine. In fact, while the recall is higher than the precision for Scholarcy, the contrary is true for Cermine. A huge contribution to the Scholarcy bad performance in the precision level is made by its mismatches in the identification of the first bibliographic reference in the paper. Indeed, the number of references identified by Scholarcy is normally superior to the correct number of references proportionally with respect to the correct number of references. For instance, AGR-BIO-SCI_1 has 34 references and the Scholarcy output 51, instead ART-HUM_3 has 99 references but the output has 131 references, and finally HEA-PRO_30 has 11 references and 19 are retrieved. The number of references is so high since some parts of the text are identified and wrongly tagged as references. Nonetheless, the recall is pretty high, which means that, of the references correctly extracted a high number is correctly tagged.

For what concerns the second cluster, ExCite and GROBID, are the tools whose values fall into it. Their f-score values for the references extraction are close to 0.55. By comparing their results, it comes out that GROBID has a higher recall value while ExCite presents a higher precision. Also, while the precision and recall values of GROBID are really close, ExCite has a considerably better performance in the precision rather than in the recall. These results show that the medium number of references extracted by ExCite is lower than the number of actually correct ones.

As regard the remaining two clusters, these are identifiable with single tools, Science Parse and Pdfssa4met. Science Parse, indeed, is in the third cluster, between 0.3 and 0.4. Indeed, the quality of the references extraction and parsing is low since it meanly identifies less references than the ones required, as it can be derived by the lower value of the recall with respect to the precision. Also, one of the main flows visible in the Science Parse results is the fact that in many cases it is not able to separate the venue from the other journal values, e.g. volume and issue, providing results like this one: `<title>Applied Animal Behaviour Science 143, 9ÔÇô17.</title>`. Also, the fact that for this research it has been used an older version than the last one published by the tool creator may have negatively affected the results.

Finally, Pdfssa4met presents the lowest score with values close to zero (0.1 or less). Surely, the fact that the tool is not up to date provides a valid reason to explain such negative results. *Figure 14* shows the results of this task compared in a bar chart, which provides a visual representation of the levels in which the tools results are located. Also, while showing the levels of the f-score as image 1, it allows to see the distance which in each tool encompasses between the precision and the recall, really close in tools like Anystyle and Cermine, but pretty high instead for Scholarcy and Science Parse.

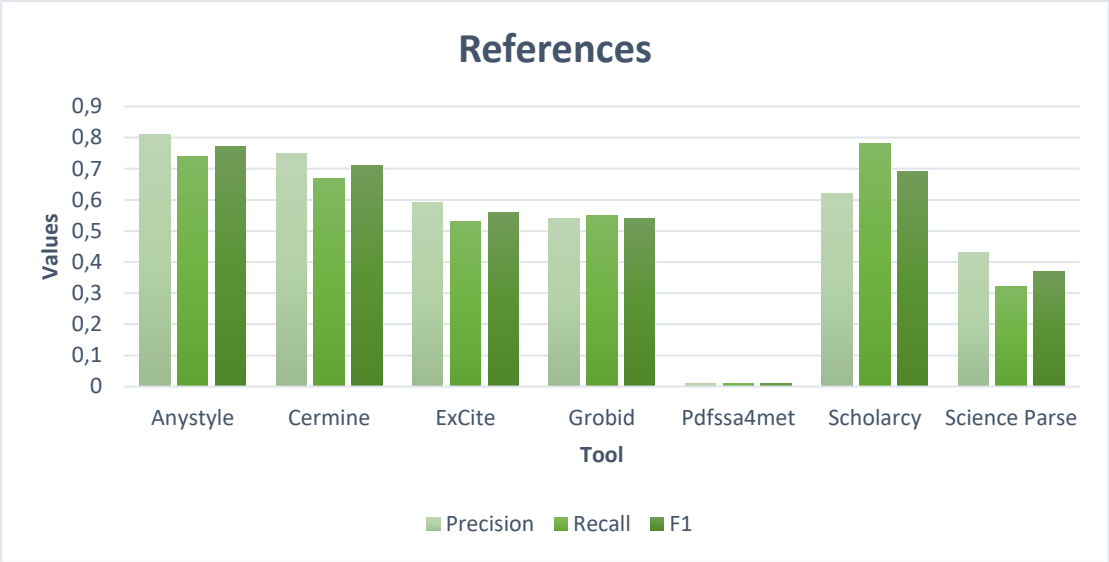

Figure 14. Barchart with precision, recall and f-score obtained in the references extraction.

As concerns the results obtained by the tools at the metadata level, *Figure 15*, offers a different scenario with respect to the references one. Indeed, this part focuses on the metadata which have been recognized by the tools. Indeed, these values represent the analysis at the level of the metadata inside the references identified as correct. Which means that, at this level, the focus is on the quality with which the tool has been able to identify the single elements contained by the references. This level represents the metadata identifiable in the intersection between all the references identified by the tool and the actually correct references. In this way it is possible to verify the quality of the tools in identifying the correct elements inside the correct bibliographic reference. An issue to report, as regards the evaluation of this level, is related to the extension of the elements set with which the tools are able to tag the references components. As seen in the previous chapters, the number of elements identifiable by each tool is variable between a few basic information to more than twenty tags. In this context having more or less metadata is relevant since, on the one hand, having too few metadata identifiable doesn't allow a proper recognition of the full set of necessary metadata. It is the case of common metadata, e.g., notes or publication place which are not identified by various tools, or of more specific information, e.g., the identification of the single elements composing the personal names. For instance, while GROBID can distinguish between forename, surname, and particles, Science Parse and Scholarcy only identify a generic personal name, including forename, surname etc. Because of the missing identification of the names' subparts, the value of the metadata identified is lower since a higher value is guaranteed to the tools which distinguish the parts of the name. Normally, since only the surname and the first name are considered, two points are given in case these two metadata are identified, one point otherwise. Vice versa, if a tool is able to recognize many

different fields, the risk is to mislead the data in the references and attribute the wrong metadata to the bibliographic entry parts.

The results provided in *Table 31*, in a certain way, provide an answer to this issue. Indeed, the tools which is concretely less able to identify a consistent number of metadata, Pdfssa4met, has the lowest score among the tools. Science Parse follows it both for number of metadata identifiable and for f-score in the metadata extraction task. The following tool is Scholarcy which reaches the 0.81, while all the other tools have high scores, above 0.9. All the tools with a value above the 0.8 have at least twelve metadata identifiable. Thus, it is observable a better performance of the tools able to capture a high number of metadata, even if it does not guarantee the highest score, and the risk of misinterpretation of the contents is not concretely relevant.

Table 31. Results of the metadata for the different tools.

	Precision	Recall	F1
Anystyle	0,93	0,97	0,95
Cermine	0,94	0,94	0,94
ExCite	0,93	0,92	0,92
Grobid	0,86	0,97	0,91
Pdfssa4met	0,14	0,29	0,19
Scholarcy	0,96	0,70	0,81
Science Parse	1,0	0,55	0,71

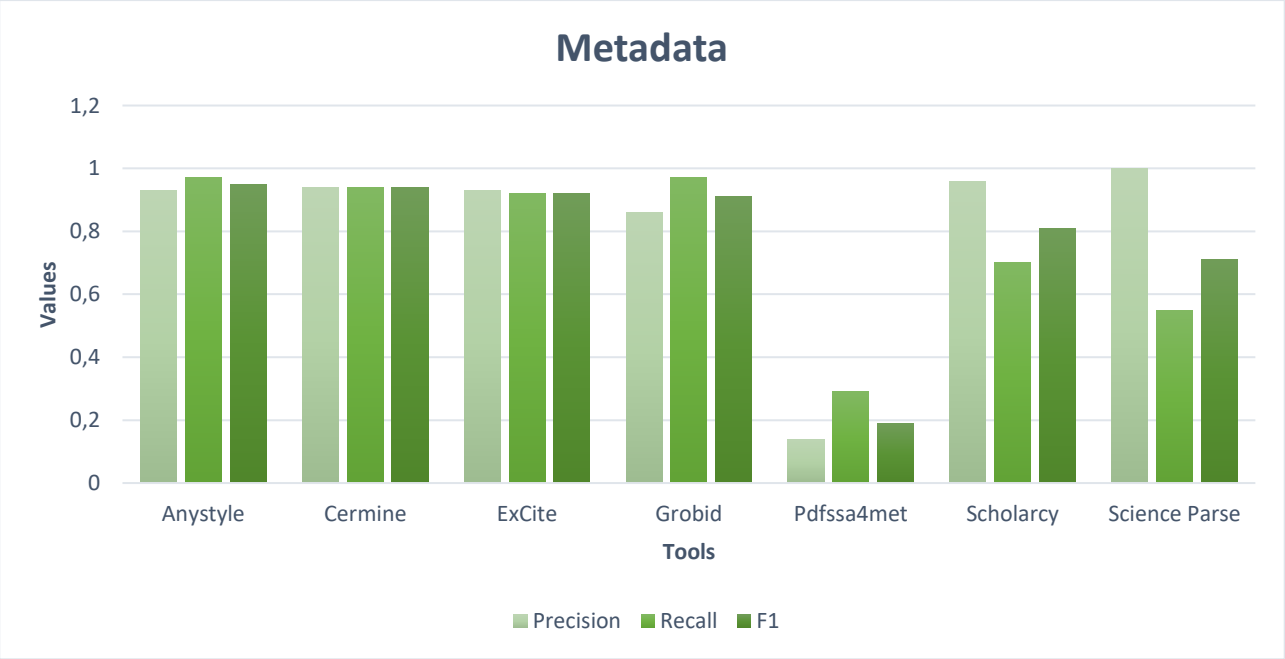

Figure 15. Barchart with precision, recall and f-score obtained in the metadata extraction.

Finally, the analysis on the references presents some further insights in the tools working. Again, Anystyle has the best f-score among the tools, followed by Cermine and GROBID, both with 0.86. Between these three tools, and as consequence between all the tools, is GROBID the one which obtains the best score in the recall, thus the one identifying the highest number of correct contents. The top precision, instead, is obtained by Anystyle with a score of 0.86. Lower scores are obtained by ExCite and Scholarcy, which nonetheless lie above 0.75. Science Parse instead, presents the highest precision among all the tools but the f-score is reduced by the recall. This difference can be explained with the fact that only the contents of the correct metadata identified in the correct references are considered.

Moreover, all the metadata can be identified only once, apart from the authors. Thus, the number of correctly identified metadata is necessarily high since, only the fact that a reference is recognized means that at least one out of four metadata identifiable has been identified. Nonetheless, such a high precision means that the identification of the remaining metadata (which in most cases coincide with the authors) is high and that the number of wrongly identified is low. At the same time, the fact that only so few metadata can be identified lowers the number of correctly identified metadata since many of the correct ones cannot be identified at all. Finally, Pdfssa4met closes its series of results confirming its poor quality in the extraction tasks.

Figure 16 graphically shows the similarity of precision and recall values characterizing Anystyle, Cermine and ExCite. At the same time, it shows the fact that GROBID obtains the highest recall and Science Parse the highest precision, while pdfssa4met present, again, the lowest score among the tools. Finally, it is shown the relevant distance between precision and recall in Scholarcy and Science Parse.

Table 32. Results of the contents for the different tools.

	Precision	Recall	F1
Anystyle	0,87	0,91	0,89
Cermine	0,86	0,87	0,86
ExCite	0,79	0,79	0,79
Grobid	0,81	0,92	0,86
Pdfssa4met	0,07	0,14	0,09
Scholarcy	0,90	0,65	0,75
Science Parse	0,94	0,51	0,66

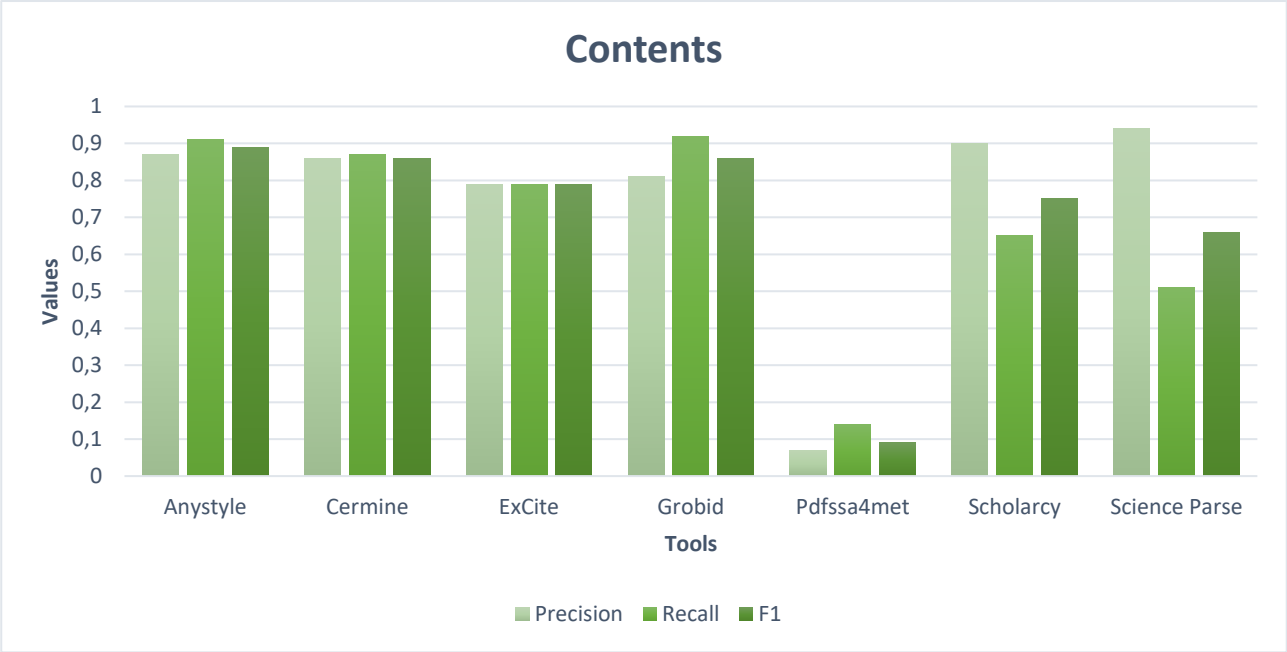

Figure 16. Comparison of the results on the contents

6.2 Influence of the Research Fields on the Results

One aspect requiring a particular attention regards the results obtained by the selected tools in each single research fields. Indeed, the analysis on the single components of the dataset may provide a deeper view in the extraction and parsing quality of the tools. *Figure 17* portraits a complex picture of the references extraction results at the research fields level. As it can be seen in this picture, Pdfssa4met did not make a good performance in any of the fields, instead, its results are close to zero in all the fields. Differently, all the other tools show a high variance in the identification quality with respect to the references, depending on the single fields. In some cases, a trend guiding the tool is visible, for instance the levels of Anystyle are mediumly closer to 0.8, the GROBID ones to 0.4 and the Science Parse ones to 0.2. Nonetheless, we can identify, on the one hand, fields in which the tools have generically better (or worse) performances and, on the other hand, fields in which one or two tools have better performances with respect to the others. In support of the first typology, we can see that ART-HUM, BUS-MAN-ACC and PSY have values lying above (or really close to) the 0.6 in all the fields, exception made for Pdfssa4met. A similar path can be observed for AGR-BIO-SCI with the only exception of Science Parse. Differently, ENE, MAT-SCI, MUL and PHY-AST report bad scores for almost all the tools involved. Indeed, in all those cases, the maximum score obtained arrives to 0.7 but the mean is closer to 0.4. In other cases, things are really variable and the difference between the tools is highly visible. For instance, Anystyle outperforms the other tools in the fields of DEC-SCI, ECO-ECO-FIN, SOC-SCI and, noticeably, the Z-NOTES-TESTS. Cermine has the best

performance in BIO-GEN-MOL, CHE-ENG, MED and VET. ExCite, has a better performance in ENV-SCI and NUR, with values close to Anystyle e Scholarcy. Instead, it has an extremely bad performance in the VET field. This fact is due to two factors working contemporaneously and showing a problematic aspect of the tool. What happens is that in VET_54 only one reference is extracted by ExCite and it does not match any of the correct references in the gold standard. At the same time, in VET_53, even if the references are for the most correctly extracted ExCite reveals a systematic issue in the distinction between the journal articles dates and volume numbers. Indeed, these two values, in all the references identified are identified and tagged as one. Because of these two different reasons, the score in this field is 0. GROBID, on the one hand, does not obtain the best score in any of the fields, but, on the other, it does not present drastically low values either. Scholarcy, instead, reports the best score in IMM-MIC and CHEM, with a score close to the one obtained by Anystyle. Finally, Science Parse does not obtain neither the best score in some fields nor is close to good results in any of the fields. The number of references extracted, even if distant from the Pdfssa4met results, is nonetheless really low, with the only exceptions of the fields whose references are identified as generically easily to parse.

From the results it is possible to see that three identifiable fields typologies with respect to the results obtained by the tools:

- *Fields which are difficult to analyse for nearly all the tools*, with different scores depending on the single tools. The reasons behind this difficulty may lie on the organization of the references in the journals publishing in those research fields. It may be, for instance, the layout, which metadata are required or the citation style. Another reason could be identified in the typologies of resources cited. For instance, the strong presence of publications different from the journal articles, the most widely identifiable publication type[57], may provide a low score to the extraction task.
- *Fields easy to parse for almost all the tools*. The papers published in these fields probably present the references structured in such a way, on the basis of the same parameters described in the previous point, that allows all the tools to parse them efficiently, regardless of their functioning.
- *Fields well identifiable only by some tools*. This is the most widespread case in the current dataset. In each single field the best scores are obtained by a minimum of one and a maximum

[57] This statement refers to the fact that some of the tools are able to extract only the metadata of the journal articles or that, even if the tools may be able to identify different resourcs concretely they are not able to do it, for instance Cermine which is not able to correctly identify the book chapters metadata.

of three tools. In this case the reasons may lie either on the functioning on the tool or the pagination of the references in the paper.

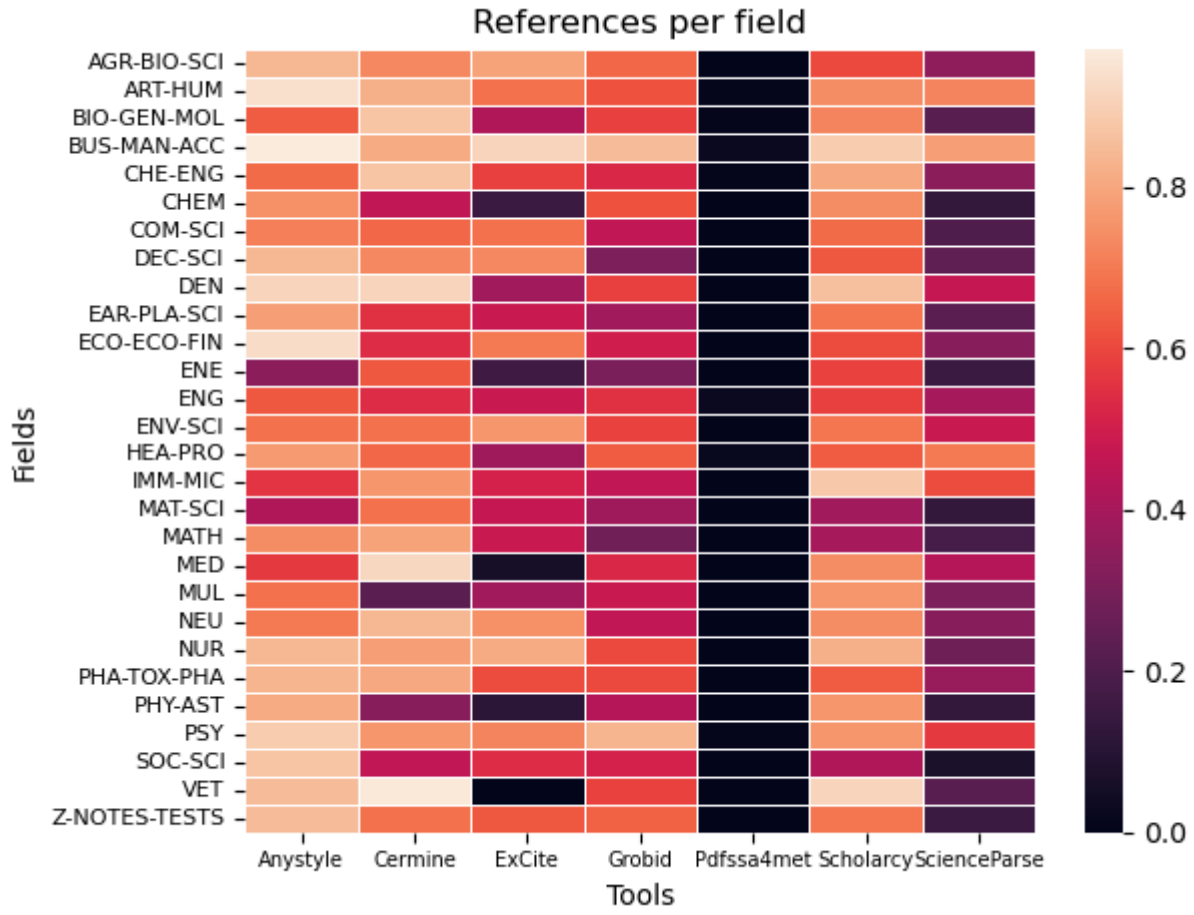

Figure 17. Comparison of the F-SCORE results per fields on the references.

The situation of the metadata is radically different from the references one. Indeed, as it is shown in *Figure 18*, the metadata are identified with a really high precision by almost all the tools with only a few exceptions. What comes out is that Anystyle, Cermine, ExCite and GROBID have their minimum result close to 0.7 with only two exceptions, in the ENE field for Anystyle and in VET for ExCite. Pdfssa4met has again the worse results even if, with respect to the references extraction, in five research fields the results are higher than 0. Finally, Scholarcy and Science Parse have mediumly lower values with respect to the previous tools, but nonetheless they are stretched around 0.7-0.8. This aspect can be traced back to the fact that the reason behind the general goodness of the results with respect to the references may lie in the fact that are taken into consideration only the metadata of the identified references. Nonetheless, the goodness of those results shows that the tools ability in identifying the metadata inside the identified references is optimal. Also, the aim of this analysis was to verify whether the fields had any influence on the tools in the task of identifying the references

metadata. Through the observation of the results shown in *Figure 18*, it comes out that the fields seem to have no particular effects on the identification of the metadata.

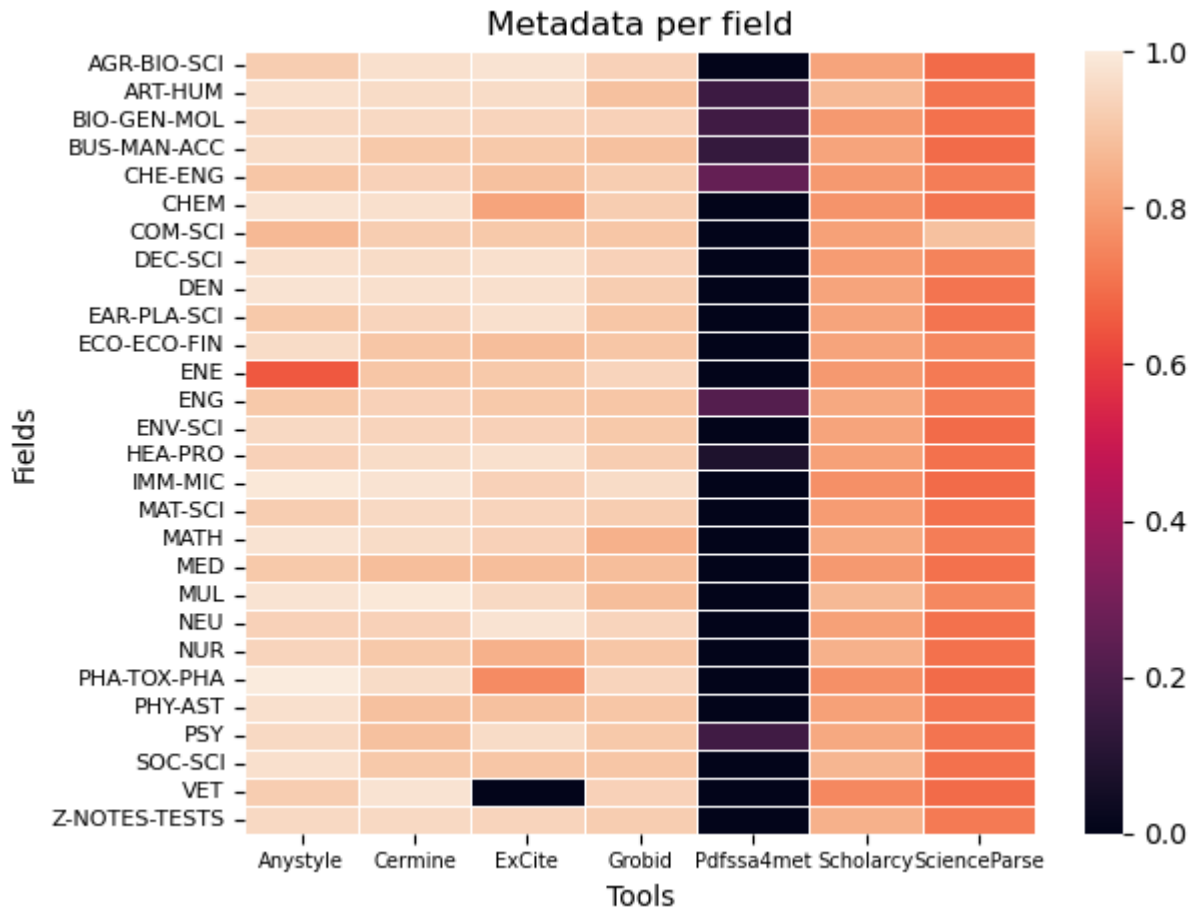

Figure 18. Comparison of the F-SCORE results of the metadata for the fields.

Finally, the contents show an even different scenario with respect to the previous two levels. Indeed, if the fields strongly influenced the references and the metadata were optimally identified by all the tools, with the due exceptions, the citations represent a synthesis of these two aspects. Indeed, while the results maintain a generically high score, the fields seem to influence the medium results. Indeed, on the one hand it is clear that, once arrived at this point, we already know that at least a number of metadata between one and three are surely correct since the references are considered correct on this basis. At the same time, we already know that the metadata have been identified with a high level of precision. Thus, the results obtained at this point regard the level of correctness on the texts of the metadata correctly identified in the correct references. The same general trends identified in the results of the metadata extraction can be viewed also at this stage, but with relevant differences in the single cases. For instance, Anystyle presents values which are around 0.8, while in the metadata level they were around 0.9. Also, it presents two extremely low values, against the one reported in *Figure 18* for the metadata level. A similar scenario is reported for Cermine and Grobid, for which the mean

values are slightly lower with respect to the metadata ones, but no particularly low values have been registered. Instead, ExCite reports a completely different scenario with respect to the metadata one. Indeed, in some specific fields the results are drastically lower with respect to the quantity of metadata correctly identified. In some cases, these drawbacks can be associated with the references ones: BIO-GEN-MOL, CHE-ENG, CHEM and PHY-AST. In this case it is clear that the fields negatively affect the results of the references extraction task. Indeed, even if the metadata are correctly identified, the same contents which were not well identified in order to let the references be accepted are now negatively affecting the contents metadata of the non-necessary metadata. Pdfssa4met, instead, maintains a high record for three out of the five fields identified with a minimally relevant value. The relevant thing to notice is that the fields where these good results (with respect to the mean values of the tools) are obtained in the same fields where ExCite obtains low values. Finally, Scholarcy and Science Parse maintain a trend similar to the metadata one, even if with a general slightly diminishing in the values.

To conclude one last observation can be made on the results on the fields. Even if not at the same levels as for the references, the fields seem to have a certain influence on the tools results. As previously mentioned, the same fields that provided a negative result to Cermine have the same effects on the metadata contents. Differently some fields seem to have a negative effect on more than one tool: EAR-PLA-SCI counts the highest number of low values among the tools. Instead, DEN, MUL and SOC-SCI contents are the ones which are best identified by the tools (considering the mean values of each tool). One interesting aspect about these results is the fact that the ones which obtained the best score in the contents evaluation obtained bad results in the references, but the contrary is not true. One possible explanation is that while in some disciplines the errors are specific to some references and, thus, excluded in the phase of the references selection, in some other disciplines the errors are systematic, either for all the parsers or only for some of them, and not only related to wrongly identified references. An exemplificative case is reported in CHEM_12 parsed by ExCite. Indeed, the first case represent a systematic error which provide a bad f-score in the references extraction and also in the content extraction. The starting point is the fact that the references do not include the article title but only the journal, volume and pages are reported. As consequence, ExCite is not able to parse well the references and provides two or three references merged in one and the related metadata are sometimes confused (e.g. issue and volume are usually inverted). Thus, when it comes the moment of identifying the reference, only one of them is identified in the metadata of the output references, the first one whose metadata are compatible with the ones reported in the output reference. This procedure is due to the fact that no references can contain two references from a logical point of view, thus only one reference can be recognized in another one on the basis of the

selected fundamental metadata. Thus, while the references register a low score because only a few references are correctly identified and the inner metadata have a low precision, the contents have a low score because the metadata are wrongly identified inside the reference.

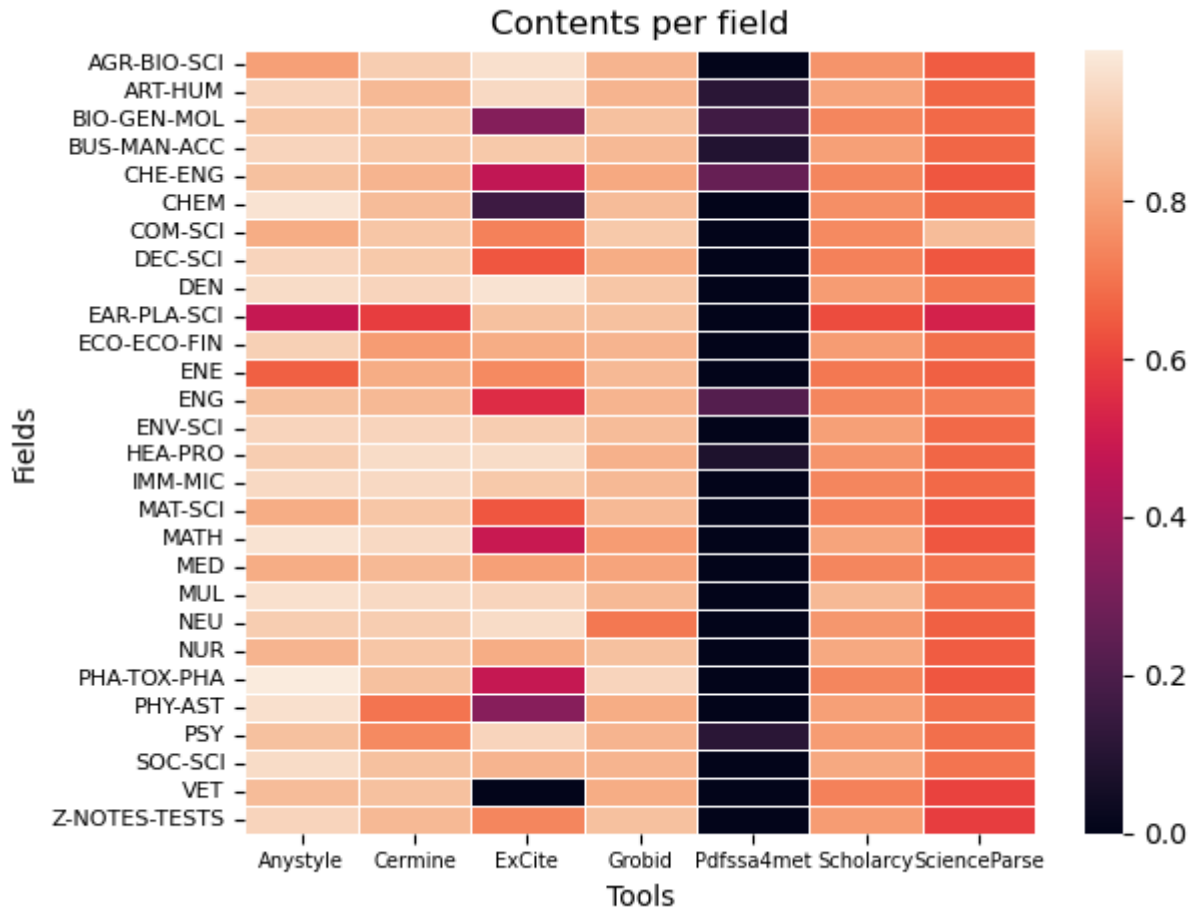

Figure 19. F-SCORE results of the content extraction per field.

6.3 Influence of the Layout on the Results

One last aspect that presents interesting features to investigate is the distribution of the results among the different types of layouts. Indeed, investigating this aspect, even if in small scale can provide more specific insights into the tool capabilities with respect to the references pagination. The focus of this analysis is on checking whether the tools had better performance on one column layout with respect to the two or three columns' ones, or vice versa. Indeed, by looking at those papers whose references in some cases were not parsed at all, it comes to sight that none of them are paginated in one column. Thus, further research has been carried out in order to verify whether this difference is systematic or if concretely the distance is not that much accentuated. One thing to anticipate is that

only three examples of three columns layout papers were present in the dataset, thus they are less represented of the two columns (32) and one column one (21). The results, nonetheless, are pretty interesting. The trend of the tools can be observed in *Figure 20*, where the values obtained by the tools are compared on the three layout types.

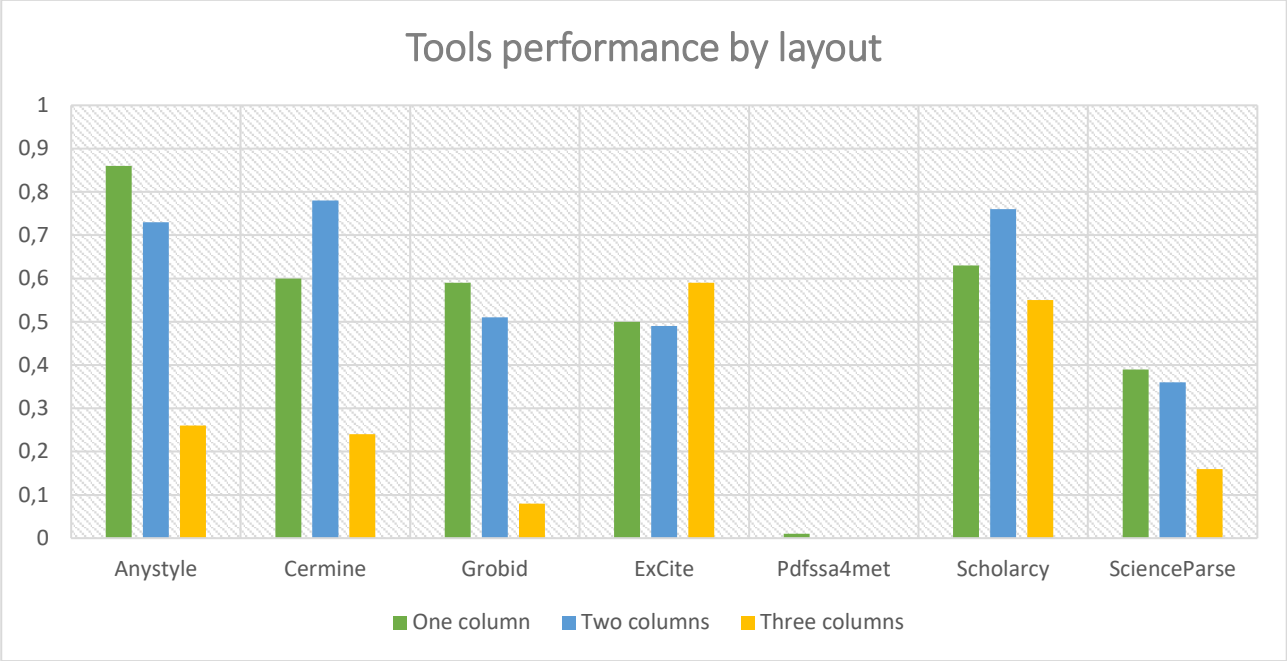

Figure 20. Tools performance on the basis of the layout

It comes out the fact that, for all the tools, the less correctly identified layout typology is the three-columned one. Also, on average, it is recognized with a significantly lower score with respect to the other two layout types. The only tool for which this trend is inverted, also with a noteworthy difference, is ExCite. Indeed, the number of identified references in this papers typology outperforms the one- and two-columned papers, which, instead, obtain similar scores. As regards the remaining tools, in four out of six of them the one columned layout obtains the best performance, while in the remaining two, Cermine and Scholarcy, is the two columned layout to outperform the others results. From this observation derives the confirm that the references formatted in one column are more easily identifiable by most of the tools. Nonetheless, there is not enough distance from the two columns layout results to affirm that the performance is far superior. Indeed, as it can be observed in the graphic 8, even if the absolutely best performance is obtained on the one column layout papers, the values obtained on the two columns layout are mediumly higher in the range of values 0.7-0.8 or, at least, in the same range as the one column ones, i.e. 0.3 – 0.6. Thus, we can affirm that, in general, the tools are able to carry out a slightly better performances on the papers where the bibliographic references are formatted in one column. This value is followed by the two columns layout, with a slightly worse

performance. Finally, the worst scores are obtained by the references structured in three columns, which in some cases are not even identified.

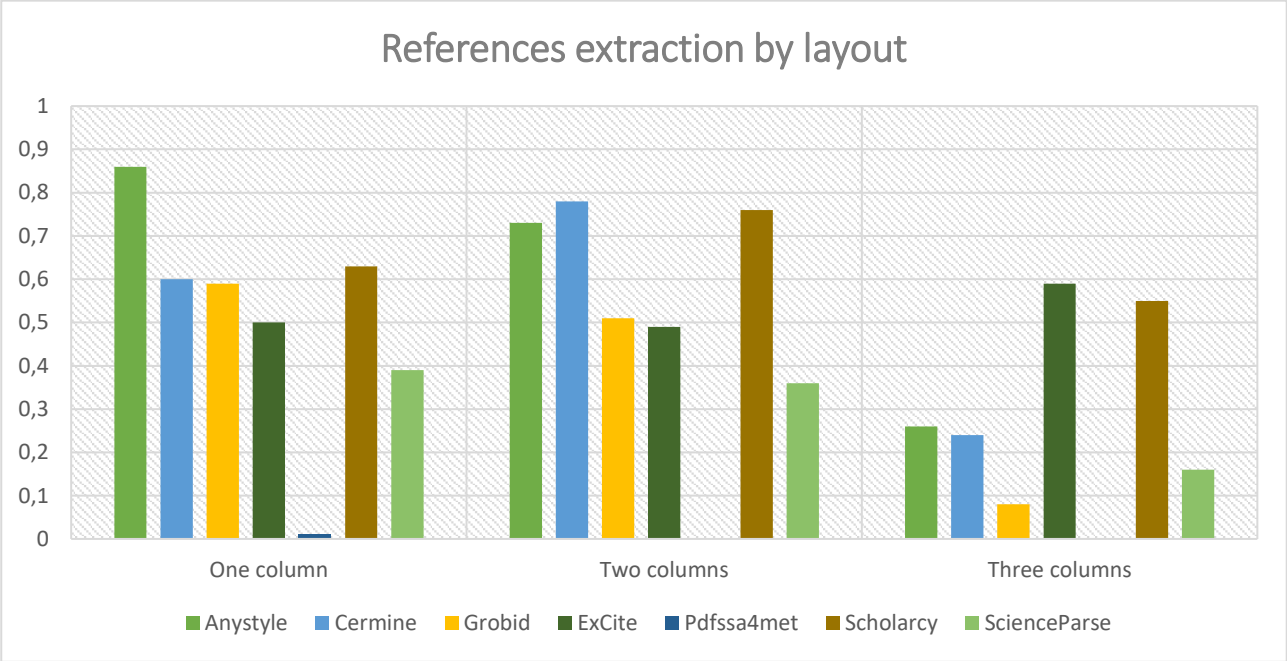

Figure 21. Tools performance on the basis of the layout. On the x axis the layout typologies

6.4 Discussion in a Nutshell

To conclude, we can affirm that there is one tool which outperforms the others in all the three tasks proposed (i.e., references, metadata and contents), Anystyle. Anystyle obtained the best f-score in all the extraction levels analysed and that its overall results are impressive. Nonetheless, the analysis of specific subtasks has shown further insights that can be interesting from a concrete perspective. Indeed, it is clear that both the other two investigated factors, the research field and the paper layout, influence the tools single scores. Indeed, Anystyle is outdone in all the three analysis levels by all the other tools in fields like ENE, where its score is really low. In other cases, it is outperformed by other tools only in the task of references extraction like in the EAR-PLA-SCI field or only in the references extraction, like in MED. At the same time the analysis on the layout showed that while obtaining the highest score in the results of the one column paginated references, Cermine and Scholarcy had a slightly better performance in the two-column layout and ExCite and Scholarcy were definitely better in extracting the references from the three columns layout. Indeed, it is not a case that Cermine and Scholarcy had the best results after Anystyle, with values close to the Anystyle ones. Instead, for what

concerns the metadata content, the tools performing similar results to Anystyle are again Cermine and GROBID, which has the best score for the metadata contents recall.

In conclusion, we can affirm that Anystyle is the tool which can extract and parse the bibliographic references from full-texts PDF papers obtaining the overall best score. Nonetheless, according to the specific peculiarities of each subtask, it may be combined with other tools in order to work on the specific subtasks they are good at, with the aim of obtaining a final better performance.

7. Conclusions

The aim of this thesis was to retrieve from the available literature all the tools able to extract the bibliographic references from full text PDF papers. The research questions which have guided this work were related, first of all, to the identification of these tools, then, to their evaluation in order to verify whether one of them works better than the others, if the single fields influence the tools performance and, finally, understanding whether the tools have better or worse performances on the basis of the papers layout features. In order to address these questions different steps have been carried out.

The first step of this research consisted in performing a systematic literature review, at the end of which a number of tools was selected. To the tools identified some further criteria have been applied, in order to exclude the non-valid tools. In the end only seven tools have been selected: Anystyle, Cermine, ExCite, GROBID, Pdfssa4met, Scholarcy and Science Parse. This answers the first question, regarding which tools carry out the bibliographic references extraction. Then, a dataset of papers selected before the beginning of the research from different research fields is applied as input to the parsers. At this point the extraction and parsing tasks are carried out by the tools which return as output one file for each input file containing the references that it was able to identify in a structured format. Then, the results obtained from the extraction and parsing phases are converted to the language of the gold standard references, compared and evaluated. The focus of the comparison has been directed to three levels inside the references: the correctly extracted references, the metadata correctly extracted in the correctly identified references and, finally, the contents of the correctly identified metadata. The evaluation has been carried out by computing precision, recall and f-score for each of the selected levels of analysis. At this point the answer to the second question was provided. Indeed, it comes out that Anystyle, a CRF based tool, outperforms the other tools in all the three levels of analysis.

Nonetheless, two other aspects of the results have been investigated in order to answer the second part of the second question, regarding the role of the single fields in the tools parsing performance and the influence of the papers layout in the extraction task. As regards the former analysis, the results for the single papers have been combined to provide the results per fields. The results of this task show that in some cases Anystyle is outperformed by other tools. These cases are either isolated, when one tool outperforms it in one task, or systematic, in case Anystyle obtains particularly bad results. The second analysis, about the layout reorganized the results in order to show the level of correctness among the paper formatted in one, two or three columns. The results show that the papers

with one column are the best identified and that Anystyle has the best performance in this case. Instead, in case the references are formatted in two columns the best score is obtained by Cermine, and in three columns by ExCite. Thus, even if, among the seven selected tools, Anystyle has outperformed the others in all the three analysis levels, nonetheless, from the investigation of the two other tasks it comes out that under some specific conditions other tools can outperform the results of Anystyle. It is the case of Cermine with papers with the references formatted in two columns or ExCite for the three columns. Or, for instance, in case the research field is Immunology Microbiology, Scholarcy outperforms Anystyle. Thus, while the best solution for the task of bibliographic references extraction and parsing, is Anystyle, a cooperation between the tools on the basis of the specific subtasks may be relevant in order to obtain the best result possible from all the tools.

This work presents three major limits. First of all, the dataset selected as input is quite small in size with respect to the average dataset contents required for this kind of study. Indeed, even if providing a large number of research fields, each of them was provided with only two fields, which is enough to provide an insight but not to have definitive views on the topic. In second place, the tools have been used in a way out-of-the-box, without any training. This lack of training, in particular for the CRF based tools may have had as consequence a loss in performance. Indeed, as it has been shown in a recent work (Tkaczyk et al. 2018b) when the tools are retrained the performance of the tools obtains better scores. Finally, as concerns the comparison of the output data against the gold standard one, it has been selected the Levenshtein distance as only metrics to measure the distance between all the metadata. Nonetheless, other measures have been identified to outperform the Levenshtein as the best measure to compute the similarity between two names in text retrieval tasks. Indeed, studies report that the best solution for the named entities matching is the hybrid method named soft TF-IDF (Cohen, Ravikumar, and Fienberg 2003). This metric is reported to outperform the other methods in all the tasks proposed. Thus, it seems the best solution in order to compute the similarity in this set of strings. In particular the hybrid methods are suggested for this kind of task, where short strings are compared. Thus, in further studies, it would be preferrable to use this measure instead of the more generic Levenshtein.

Future works may take into consideration more tools than the ones accepted in this one, by raising the acceptability threshold and excluding some of the parameters adopted here. Indeed, this work was directed to a specific target, which required the exclusion of more complex to use tools. But changing the target may allow to have more tools to study and compare in future works. For instance, allowing to include also trainable models could provide more insights into the potentiality inherent the field and not studied in this thesis. Also, in a future development of this research it will be useful to provide

a more complete dataset in order to consolidate the results obtained in this research. Indeed, working with a higher number of input data may allow to recognize aspects such as the effective quality of the parsers in the identification of the references formatted in three columns, since in this dataset their were underrepresented. Being available a valid dataset to use as input, it would be possible to verify whether the results obtained in this research reflect the actual trend or if the underrepresentation has brought to a partial interpretation of it.

Bibliografia

Ammar, Waleed, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, et al. 2018. 'Construction of the Literature Graph in Semantic Scholar'. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, 84–91. New Orleans - Louisiana: Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-3011.

Azimjonov, Jahongir, and Jumabek Alikhanov. 2018. 'Rule Based Metadata Extraction Framework from Academic Articles'. *ArXiv:1807.09009 [Cs]*, July. http://arxiv.org/abs/1807.09009.

Bast, Hannah, and Claudius Korzen. 2013. 'The Icecite Research Paper Management System'. In *WISE 2013: Web Information Systems Engineering – WISE 2013*, 396–409. https://doi.org/10.1007%2F978-3-642-41154-0_30.

Bhardwaj, Akansha, Dominik Mercier, Andreas Dengel, and Sheraz Ahmed. 2017. 'DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction'. In *Neural Information Processing*, edited by Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy, 10635:286–93. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-70096-0_30.

Biblio Citation Parser. 2004. https://metacpan.org/release/MJEWELL/Biblio-Citation-Parser-1.10.

BibPro. n.d. https://github.com/ice91/BibPro.

Bilbo2 (version 1.2.0). n.d. https://github.com/OpenEdition/bilbo2.

Blumberg, Robert, and Shaku Atre. "The Problem with Unstructured Data." *DM Review Magazine*, February 2003, 42–46. http://www.dmreview.com/article sub.cfm?articleId=6287.

Brownlee, Jason. 2017. 'How to Clean Text for Machine Learning with Python'. *Machine Learning Mastery*, 18 October 2017, sec. Deep Learning for Natural Language Processing. https://machinelearningmastery.com/clean-text-machine-learning-python/.

Cb2Bib. 2021. https://www.molspaces.com/cb2bib/doc/overview/.

CERMINE - Web Service. n.d. http://cermine.ceon.pl/cermine/index.html.

Chen, Chien-Chih, Kai-Hsiang Yang, Chuen-Liang Chen, and Jan-Ming Ho. 2012. 'BibPro: A Citation Parser Based on Sequence Alignment'. *IEEE Transactions on Knowledge and Data Engineering* 24 (2): 236–50. https://doi.org/10.1109/TKDE.2010.231.

Chen, Chien-Chih, Kai-Hsiang Yang, Hung-Yu Kao, and Jan-Ming Ho. 2008. 'BibPro: A Citation Parser Based on Sequence Alignment Techniques'. In *22nd International Conference on Advanced Information Networking and Applications - Workshops (Aina Workshops 2008)*, 1175–80. Gino-wan, Okinawa, Japan: IEEE. https://doi.org/10.1109/WAINA.2008.125.

Cioffi, Alessia. 2022a. “Code for Converting Different Formats to TEI XML and Evaluation of the Results”. Zenodo, https://doi.org/10.5281/zenodo.6182128.

Cioffi Alessia. 2022b. “Data for Testing and Evaluating References Extraction and Parsing Tools”. Zenodo. https://doi.org/10.5281/zenodo.6182066.

Cioffi, Alessia. 2022c. “Systematic Literature Review about Software for References Extraction.” protocols.io. https://dx.doi.org/10.17504/protocols.io.buz9nx96.

Citation. 2016. https://github.com/nishimuuu/citation.

Citation-Parser. n.d. https://github.com/manishbisht/Citation-Parser.

Clark, Christopher, and Santosh Divvala. 2016. ‘PDFFigures 2.0: Mining Figures from Research Papers’. In *JCDL*. https://ai2-website.s3.amazonaws.com/publications/pdf2.0.pdf.

Cohen, William W., Pradeep Ravikumar, and Stephen E. Fienberg. 2003. ‘A Comparison of String Distance Metrics for Name-Matching Tasks’. In *IIWEB’03: Proceedings of the 2003 International Conference on Information Integration on the Web*. https://dl.acm.org/doi/10.5555/3104278.3104293.

Constantin, Alexandru, Steve Pettifer, and Andrei Voronkov. 2013. ‘PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature’. In *Proceedings of the 2013 ACM Symposium on Document Engineering*, 177–80. Florence Italy: ACM. https://doi.org/10.1145/2494266.2494271.

Content ExtRactor and MINEr (version 1.13). n.d. https://doi.org/10.5281/zenodo.569829.

Councill, Isaac, C. Lee Giles, and Min-Yen Kan. 2008. ‘ParsCit: An Open-Source CRF Reference String Parsing Package’. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC ’08)*. Marrakech, Morocco. https://www.cs.brandeis.edu/~marc/misc/proceedings/lrec-2008/pdf/166_paper.pdf.

Davis Jr., Clodoveu A., and Emerson de Salles. 2007. ‘Approximate String Matching for Geographic Names and Personal Names’. In , 49–60. Campos do Jordão, Brazil: INPE.

Elsevier. 2021. *Mendeley*. https://www.mendeley.com/.

Ferrés, Daniel, Horacio Saggion, Francesco Ronzano, and Alex Bravo. 2018. ‘PDFdigest: An Adaptable Layout-Aware PDF-to-XML Textual Content Extractor for Scientific Articles’. In . https://www.aclweb.org/anthology/L18-1298.pdf.

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, et al. 2018. ‘Science of Science’. *Science* 359 (6379): eaao0185. https://doi.org/10.1126/science.aao0185.

FreeCite. 2008. https://github.com/miriam/free_cite,Miriam Goldberg.

Fuchun, Peng, and Andrew McCallum. 2004. ‘Accurate Information Extraction from Research Papers Using Conditional Random Fields’. In *NAACL*. https://www.aclweb.org/anthology/N04-1042.pdf.

Gao, Liangcai, Zhi Tang, and Xiaofan Lin. 2009. ‘CEBBIP: A Parser of Bibliographic Information in Chinese Electronic Books’. In *Proceedings of the 2009 Joint International Conference on Digital Libraries - JCDL ’09*, 73. Austin, TX, USA: ACM Press. https://doi.org/10.1145/1555400.1555412.

Ghavimi, Behnam, Wolfgang Otto, and Philipp Mayr. 2019. ‘EXmatcher: Combining Features Based on Reference Strings and Segments to Enhance Citation Matching’. *ArXiv:1906.04484 [Cs]*, June. http://arxiv.org/abs/1906.04484.

Gooch, Phil. 2021. ‘How Scholarcy Contributes to and Makes Use of Open Citations’. *Scholarcy*, 2021. https://www.scholarcy.com/how-scholarcy-contributes-to-and-makes-use-of-open-citations/.

Groeneveld, Dirk, Doug Downey, and Aria Haghighi. 2019. *Science Parse*. https://github.com/allenai/science-parse.

Guo, Zhixin, and Hai Jin. 2011. ‘Reference Metadata Extraction from Scientific Papers’. In *2011 12th International Conference on Parallel and Distributed Computing, Applications and Technologies*, 45–49. Gwangju, South Korea: IEEE. https://doi.org/10.1109/PDCAT.2011.72.

Hager, Chris. n.d. *PDFx* (version 1.4.1). https://github.com/metachris/pdfx.

Hashmi, Ahmer Maqsood, Faiza Qayyum, and Afzal Muhammad Tanvir. 2020. ‘Insights to the State-of-the-Art PDF Extraction Techniques’. In *IPSI Transactions on the Internet Research*, 16:60–67. http://ipsitransactions.org/journals/papers/tir/2020jan/p9.pdf.

Herzog, C, D Hook, and E Adie. 2018. ‘Reproducibility or Producibility? Metrics and Their Masters’. In *STI 2018 Conference Proceedings*, 685–87. Centre for Science and Technology Studies (CWTS). https://hdl.handle.net/1887/65257.

Hetzner, Erik. 2008. ‘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 ’08*, 280. Pittsburgh PA, PA, USA: ACM Press. https://doi.org/10.1145/1378889.1378937.

Holvitie, Johannes, and Ville Leppänen. 2015. ‘RefUTU: Automatic Bibliography Database Generation for Freely Formatted Reference Listings’. In *Proceedings of the 16th International Conference on Computer Systems and Technologies - CompSysTech ’15*, 176–83. Dublin, Ireland: ACM Press. https://doi.org/10.1145/2812428.2812469.

Hosseini, Azam, Behnam Ghavimi, Zeyd Boukhers, and Philipp Mayr. 2019. 'EXCITE – A Toolchain to Extract, Match and Publish Open Literature References'. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 432–33. Champaign, IL, USA: IEEE. https://doi.org/10.1109/JCDL.2019.00105.

Hsieh, Yu-Lun, Shih-Hung Liu, Ting-Hao Yang, Yu-Hsuan Chen, Yung-Chun Chang, Gladys Hsieh, Cheng-Wei Shih, Chun-Hung Lu, and Wen-Lian Hsu. 2014. 'A Frame-Based Approach for Reference Metadata Extraction'. In *Technologies and Applications of Artificial Intelligence*, edited by Shin-Ming Cheng and Min-Yuh Day, 8916:154–63. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-13987-6_15.

Hunter Library Research Guides. 2021. 'Get Published!' Western Carolina University. https://researchguides.wcu.edu/getpublished.

Huynh, Tin, and Kiem Hoang. 2010. 'GATE Framework Based Metadata Extraction from Scientific Papers'. In *2010 International Conference on Education and Management Technology*, 188–91. Cairo, Egypt: IEEE. https://doi.org/10.1109/ICEMT.2010.5657675.

Indrawati, Ariani, Ambar Yoganingrum, and Pradipta Yuwono. 2019. 'Evaluating the Quality of the Indonesian Scientific Journal References Using ParsCit, CERMINE and GROBID'. *Library Philosophy and Practice*, 2019.

'Initiative for Open Citations'. n.d. Accessed 9 February 2021. https://i4oc.org/.

'Junior.' 2008. In *West's Encyclopedia of American Law*, 2nd ed. https://legal-dictionary.thefreedictionary.com/junior.

Kan, Min-Yen. 2017. *SectLabel*. https://github.com/knmnyn/ParsCit/tree/master/bin/sectLabel.

Keil, Silvester. n.d. *AnyStyle-Cli*. https://github.com/inukshuk/anystyle-cli.

Keil, Silvester, Phil Gooch, Carlos Peña, Alex Fenton, Lars Willighagen, and namyra. n.d. *AnyStyle*. https://github.com/inukshuk/anystyle.

Khabsa, Madian, and C. Lee Giles. 2014. 'The Number of Scholarly Documents on the Public Web'. Edited by Ren Zhang. *PLoS ONE* 9 (5): e93949. https://doi.org/10.1371/journal.pone.0093949.

Khalid, Irfan Alghani. 2020. 'Cleaning Text Data with Python'. *Towards Data Science*, 12 October 2020. https://towardsdatascience.com/cleaning-text-data-with-python-b69b47b97b76.

Kim, Kihong, and Yeonok Chung. 2018. 'Overview of Journal Metrics'. *Science Editing* 5 (1): 16–20. https://doi.org/10.6087/kcse.112.

Kim, Young-Min, Patrice Bellot, Jade Tavernier, Elodie Faath, and Marin Dacos. 2012. 'Evaluation of BILBO Reference Parsing in Digital Humanities via a Comparison of Different Tools'. In

Proceedings of the 2012 ACM Symposium on Document Engineering - DocEng '12, 209. Paris, France: ACM Press. https://doi.org/10.1145/2361354.2361400.

King, Donald, Denis Jérome, Mary Van Allen, Peter Shepherd, and Johan Bollen. 2009. 'Tools and Metrics: Keynote Speech'. *Information Services & Use* 28 (3–4): 215–28. https://doi.org/10.3233/ISU-2008-0579.

Kitchenham, Barbara. "Procedures for Performing Systematic Reviews," July 2004. https://www.researchgate.net/publication/228756057_Procedures_for_Performing_Systematic_Reviews.

Kluegl, Peter, Andreas Hotho, and Frank Puppe. 2010. 'Local Adaptive Extraction of References'. In *KI 2010: Advances in Artificial Intelligence*, edited by Rüdiger Dillmann, Jürgen Beyerer, Uwe D. Hanebeck, and Tanja Schultz, 6359:40–47. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-16111-7_4.

Körner, Martin, Behnam Ghavimi, Philipp Mayr, Heinrich Hartmann, and Steffen Staab. 2017. 'Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications'. In *New Trends in Databases and Information Systems*, edited by Mārīte Kirikova, Kjetil Nørvåg, George A. Papadopoulos, Johann Gamper, Robert Wrembel, Jérôme Darmont, and Stefano Rizzi, 767:137–45. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-67162-8_15.

Kunnas, Elias. 2013. *PDF Structure and Syntactic Analysis for Metadata Extraction and Tagging (Pdfssa4met)*. https://github.com/eliask/pdfssa4met.

Lecy, Jesse D., and Kate E. Beatty. "Representative Literature Reviews Using Constrained Snowball Sampling and Citation Network Analysis." *SSRN Electronic Journal*, 2012. https://doi.org/10.2139/ssrn.1992601.

Levene, M. 2010. *An Introduction to Search Engines and Web Navigation*. 2nd ed. Hoboken, N.J: John Wiley.

Lopez, Patrice. 2009. 'GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications'. In *Research and Advanced Technology for Digital Libraries*, edited by Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas, 5714:473–74. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04346-8_62.

Lopez, Patrice. 2021. *GROBID*. https://cloud.science-miner.com/grobid/.

Luong, Minh-Thang, Thuy Dung Nguyen, and Min-Yen Kan. 2012. ‘Logical Structure Recovery in Scholarly Articles with Rich Document Features’. *Multimedia Storage and Retrieval Innovations for Digital Library Systems*, 2012.

Mulberry Technologies, Inc., National Center for Biotechnology Information (NCBI), and National Library of Medicine (NLM). 2015. ‘Journal Archiving and Interchange Tag Library NISO JATS Version 1.1 (ANSI/NISO Z39.96-2015)’. http://jats.nlm.nih.gov/archiving/tag-library/1.1/index.html.

Ning, Xiaomin, Hai Jin, and Hao Wu. 2006. ‘SemreX: Towards Large-Scale Literature Information Retrieval and Browsing with Semantic Association’. In *2006 IEEE International Conference on E-Business Engineering (ICEBE’06)*, 602–9. Shanghai, China: IEEE. https://doi.org/10.1109/ICEBE.2006.87.

Nishimura, Takahiro. 2016. *Parse Citation List in Paper*. https://github.com/nishimuuu/citation.

NLTK Project. 2021. ‘NLTK’. https://www.nltk.org/howto/metrics.html.

OCR++. n.d. http://www.cnergres.iitkgp.ac.in/OCR++/home/.

Ojokoh, Bolanle, Ming Zhang, and Jian Tang. 2011. ‘A Trigram Hidden Markov Model for Metadata Extraction from Heterogeneous References’. *Information Sciences* 181 (9): 1538–51. https://doi.org/10.1016/j.ins.2011.01.014.

‘P5: Guidelines for Electronic Text Encoding and Interchange’. 2021. https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

Patrice, Lopez. 2015. ‘GROBID from PDF to Structured Documents’. https://grobid.readthedocs.io/en/latest/grobid-04-2015.pdf.

Pdfextract. 2013. https://www.crossref.org/labs/pdfextract/.

Ramesh Kashyap, Abhinav, and Min-Yen Kan. 2020. ‘SciWING– A Software Toolkit for Scientific Document Processing’. In *Proceedings of the First Workshop on Scholarly Document Processing*, 113–20. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.sdp-1.13.

Reftagger. n.d. https://github.com/rmcgibbo/reftagger.

Rizvi, Syed Tahseen Raza, Andreas Dengel, and Sheraz Ahmed. 2020. ‘A Hybrid Approach and Unified Framework for Bibliographic Reference Extraction’. *IEEE Access* 8: 217231–45. https://doi.org/10.1109/ACCESS.2020.3042455.

Romary, Laurent, and Patrice Lopez. 2015. ‘GROBID - Information Extraction from Scientific Publications’. *Ercim News* 100: 41–42.

Ronzano, Francesco, and Horacio Saggion. 2015. ‘Dr. Inventor Framework: Extracting Structured Information from Scientific Publications’. In *Discovery Science*, edited by Nathalie Japkowicz

and Stan Matwin, 9356:209–20. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24282-8_18.

'Rule Based Metadata Extraction Framework from Academic Articles'. n.d.

Rusu, O., I. Halcu, O. Grigoriu, G. Neculoiu, V. Sandulescu, M. Marinescu, and V. Marinescu. 2013. 'Converting Unstructured and Semi-Structured Data into Knowledge'. In *2013 11th RoEduNet International Conference*, 1–4. Sinaia: IEEE. https://doi.org/10.1109/RoEduNet.2013.6511736.

Santos, Erika Alves dos, Silvio Peroni, and Marcos Luiz Mucheroni. "An Analysis of Citing and Referencing Habits across All Scholarly Disciplines: Approaches and Trends in Bibliographic Metadata Errors." arXiv.org, February 17, 2022. https://doi.org/10.48550/arXiv.2202.08469.

Scholarcy. 2022. 'About Us'. https://www.scholarcy.com/about-us/.

'SciWing Tutorials'. n.d. https://sciwing.readthedocs.io/en/latest/usage/tutorials.html.

Singh, Mayank. 2016. 'OCR++: A Robust Framework For Information Extraction from Scholarly Articles'. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 3390–3400. Osaka, Japan. https://www.aclweb.org/anthology/C16-1320.

Soulo, Tim. 2019. 'Ranking #1 on Google Is Overrated (Ahref's Study of 100k Words)'. *Ahref Blog*, 28 March 2019. https://ahrefs.com/blog/ranking-number-one-is-overrated/.

Suryawati, Endang, and Dwi H. Widyantoro. 2017. '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)*, 276–81. Bandung: IEEE. https://doi.org/10.1109/ICICI-BME.2017.8537772.

'The Initiative for Open Citations'. n.d. https://i4oc.org/.

Tkaczyk, Dominika, Lukasz Bolikowski, Artur Czeczko, and Krzysztof Rusek. 2012. 'A Modular Metadata Extraction System for Born-Digital Articles'. In *2012 10th IAPR International Workshop on Document Analysis Systems*, 11–16. Gold Coast, Queenslands, TBD, Australia: IEEE. https://doi.org/10.1109/DAS.2012.4.

Tkaczyk, Dominika, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018a. 'Evaluation and Comparison of Open Source Bibliographic Reference Parsers: A Business Use Case'. *ArXiv:1802.01168 [Cs]*. http://arxiv.org/abs/1802.01168.

Tkaczyk, Dominika, Andrew Collins, Paraic Sheridan, and Joeran Beel. 2018b. '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*, 99–108. Fort Worth Texas USA: ACM.

https://doi.org/10.1145/3197026.3197048.

Tkaczyk, Dominika, Pawel Szostek, Piotr Jan Dendek, Mateusz Fedoryszak, and Lukasz Bolikowski. 2014. 'CERMINE -- Automatic Extraction of Metadata and References from Scientific Literature'. In *2014 11th IAPR International Workshop on Document Analysis Systems*, 217–21. Tours, France: IEEE. https://doi.org/10.1109/DAS.2014.63.

Tkaczyk, Dominika, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. 'CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature'. *International Journal on Document Analysis and Recognition (IJDAR)* 18 (4): 317–35. https://doi.org/10.1007/s10032-015-0249-8.

'Unicode 14.0 Character Code Charts'. n.d. http://www.unicode.org/charts/.

Van Noorden, Richard. 2014. 'Global Scientific Output Doubles Every Nine Years'. *Nature News Blog*, 7 May 2014. http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html.

Verkuil, Steven. 2016. 'CITEREP - JOURNAL CITATION STATISTICS FOR LIBRARY COLLECTIONS USING DOCUMENT REFERENCE EXTRACTION TECHNIQUES'. http://purl.utwente.nl/essays/70399.

ViBRANT. n.d. *RefParse*. https://github.com/VBRANT/refparse.

Vijaymeena, M.K., and K. Kavitha. 2016. 'A Survey on Similarity Measures in Text Mining'. *Machine Learning and Applications: An International Journal (MLAIJ)*, March 2016.

WING-NUS. n.d. *Neural-ParsCit* (version 1.0.7). https://github.com/WING-NUS/Neural-ParsCit.

Wohlin, Claes. "Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering." *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, 2014. https://doi.org/10.1145/2601248.2601268.

Wu, Jian, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Tuarob, Cornelia Caragea, and C. Lee Giles. 2015. 'PDFMEF: A Multi-Entity Knowledge Extraction Framework for Scholarly Documents and Semantic Search'. In *Proceedings of the 8th International Conference on Knowledge Capture*, 1–8. Palisades NY USA: ACM. https://doi.org/10.1145/2815833.2815834.

Xiao, Yu, and Maria Watson. 2019. 'Guidance on Conducting a Systematic Literature Review'. *Journal of Planning Education and Research* 39 (1): 93–112. https://doi.org/10.1177/0739456X17723971.

Yin, Ping, Ming Zhang, ZhiHong Deng, and DongQing Yang. 2004. 'Metadata Extraction from Bibliographies Using Bigram HMM'. In *Digital Libraries: International Collaboration and*

Cross-Fertilization, edited by Zhaoneng Chen, Hsinchun Chen, Qihao Miao, Yuxi Fu, Edward Fox, and Ee-peng Lim, 3334:310–19. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30544-6_33.

Zhang, Xiaoli, Jie Zou, Daniel X Le, and George R Thoma. 2011. 'A Structural SVM Approach for Reference Parsing'. *BMC Bioinformatics* 12 (S3): S7. https://doi.org/10.1186/1471-2105-12-S3-S7.

Appendix A - Metadata Tagging and Usage

The table reports the list of all the metadata identified for the citations and the respective location inside the <biblStruct> element, all the publication types in which they are identified, the specific element which identifies it and the usage of the element.

LOCATION	METADATA	PUBLICATION TYPE	XML TEI TAG	USAGE
Analytic/ Monograph	Author	Articles, Books, Book Chapters, Book Series, Conference papers, Ebooks, Ebooks Chapter, Grey Literature, Manuals etc., Newspapers, Databases, Patents, Proceedings, Software, Standards, Technical Reports,Technical Reports Chapters, Unpublished, Web Pages, Preprints	<author> [<persName> <surname></surname> <forename [type= {"first" | "middle" | "last"}]></forename> <genName></genName> </persName>] </author>	The tag "author" identifies the authors of all the materials, either monographs (e.g. books) or articles.
Monograph	Content/ media/ carrier type (in AACR2)	Grey Literature	<note></note>	A work feature, either physical (CD) or metaphorical (thesis)
Monograph > Imprint	Date	Articles, Books, Book Chapters, Book Series, Conference papers, Ebooks, Ebooks Chapter, Grey Literature, Manuals etc., Newspapers, Databases, Patents, Proceedings, Software, Standards, Technical Reports,Technical Reports Chapters, Unpublished, Preprints	<date {when="" | from="" to=""}></date>	The dates of publication can be of different types: single or range, year or day.
Analytic	DOI	Book Series	<idno type="DOI"></idno>	A generic identifier tag with the specification of DOI.
Monograph	Editor	Articles, Books, Book Chapters, Conference papers, Ebooks, Ebooks Chapter, Grey Literature, Manuals etc., Newspapers, Databases, Patents, Proceedings,	<editor> [<persName> <surname></surname>	It identifies the work editor. In order to identify

		Software, Standards, Technical Reports, Unpublished, Web Pages, Preprints	<forename [type= {"first" | "middle" | "last"}]></forename> <genName></genName> </persName>] </editor>	editors the sigla ed./eds. should appear in the reference.
Analytic	ISBN	Book Series	<idno type="ISBN"></idno>	A generic identifier tag with the specification of ISBN.
Monograph > Imprint	Issue	Book Series	<biblScope unit="issue"></biblescope>	It identifies the issue of the book series.
Monograph	Journal title	Articles	<title level="j"></title>	Level "J" identifies journal titles.
Monograph > Imprint	Pagination	Articles, Books Chapter, Proceedings	<biblScope unit="page" from="" to="" /> or <biblScope unit="page"></biblescope>	
Monograph	Patent number	Patents	<idno type="docNumber"></idno>	A generic identifier tag with the specification of docNumber.
Monograph > Imprint	Place of publication	Books, Book Chapters, Technical Reports Chapters	<pubPlace></pubPlace>	Tag for the document publication place.
Monograph > Imprint	Publisher	Books, Book Chapters, Book Series, Grey Literature, Manuals etc., Proceedings, Technical Reports Chapters	<publisher></publisher>	Tag for the document publisher.
Series	Series Title	Book Series	<title level="s"></title>	Level "S" identifies series titles.

Monograph	Unpublished note	Unpublished	`<note></note>`	Annotation that the work has not been published yet.
Monograph	URL	Web SItes	`<ref target=""></ref>`	The URL is the link to which the resource web page.
Monograph > Imprint	Volume number (articles)	Articles	`<biblScope unit="volume"></biblScope>`	A bibleScope tag with a specification of the volume.
Series	Volume number (series)	Book Series	`<biblScope unit="volume"></biblScope>`	A bibleScope tag with a specification of the volume.
Monograph	Monographic work/Conference/Proceeding title	Books, Book Chapters, Conference papers, Ebooks, Ebooks Chapter, Grey Literature, Manuals etc., Newspapers, Databases, Patents, Proceedings, Software, Standards, Technical Reports,Technical Reports Chapters, Unpublished, Web Pages, Preprints	`<title level="m"></title>`	Level “M” identifies monograph titles.
Analytic	Work/chapter/article title	Articles, Book Chapters,Book Series, Conference papers, Ebooks Chapter, Newspapers, Proceedings, Technical Reports Chapters	`<title level="a"></title>`	Level “A” identifies article titles.