Structured references from PDF articles

Assessing the tools for bibliographic reference extraction and parsing

26th International Conference on Theory and Practice of Digital Libraries
Padua, Italy
September 22, 2022

Alessia Cioffi, Silvio Peroni*

*Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy
silvio.peroni@unibo.it – @essepuntato – https://www.unibo.it/sitoweb/silvio.peroni/en
Director of OpenCitations
silvio.peroni@opencitations.net – @opencitations – http://opencitations.net

RESEARCH CENTRE
FOR OPEN SCHOLARLY METADATA

OpenCitations

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DEPARTMENT OF CLASSICAL PHILOLOGY
AND ITALIAN STUDIES

The more literature, the more (meta)data

In past decades, the academic publishing world has needed to face an exponential increase in the volume of scientific literature materials

Providing academic knowledge and related metadata as structured & machine-readable formats revealed positive effects in the searchability and availability of such information

Big publishers have started to invest effort (and money) in the extraction and publication of their papers' metadata in structured formats, including bibliographic references (e.g. see the effort done by Initiative for Open Citations, I4OC, https://i4oc.org)

Smaller publishers have difficulties in carrying out this task independently since using and maintaining a tool (or paying a company addressing that task on behalf of the publisher) requires extra costs beyond publishers' finances

Optimal solution: adoption of off-the-shelf tools able to automatically extract and parse references from PDF files reaching a good quality

Goal

To analyse the current availability of bibliographic reference extraction tools to identify which one (used off-the-shelf, without no prior customisation and training) outperforms the others in extracting and parsing bibliographic references of academic papers

Methodology
Cioffi, A.: Systematic literature review about software for references extraction. protocols.io (2022). https://doi.org/10.17504/protocols.io.buz9nx96

Data
Cioffi, A.: Data for testing and evaluating references extraction and parsing tools. Zenodo (2022).https://doi.org/10.5281/zenodo.6182066

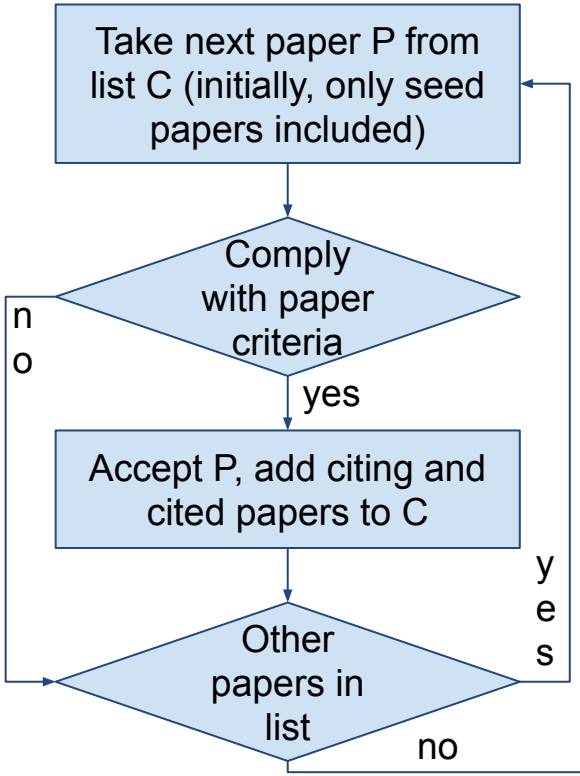
Software
Cioffi, A.: Code for converting different formats to TEI XML and evaluation of the results. Zenodo (2022).https://doi.org/10.5281/zenodo.6182128

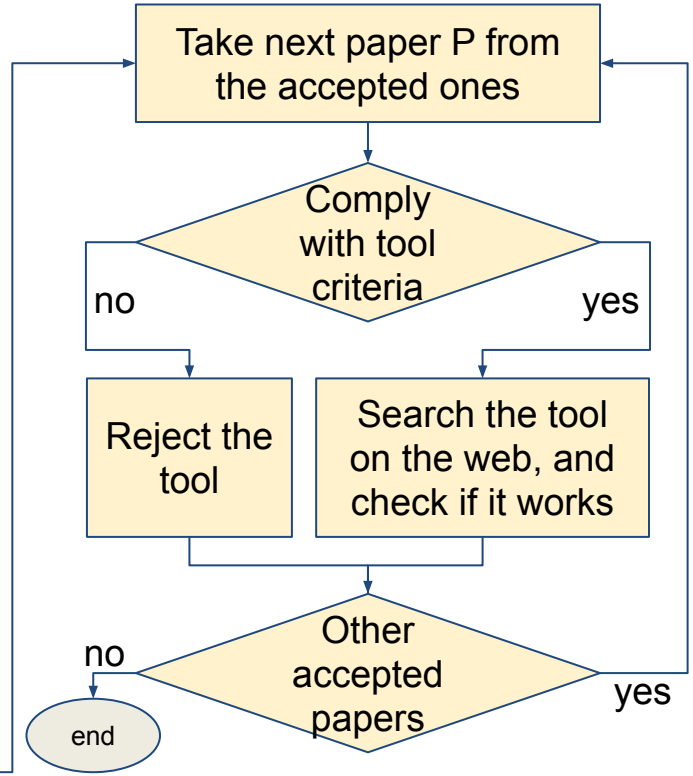
Identifying the reference extraction tools

Input:
1. Seed papers
2. Relevant words

Paper criteria:
1. Paper in English
2. Title, abstract, keywords contain relevant words
3. Paper text includes info about tools

Tool criteria:
1. Parse PDF
2. Tagging references
3. Marking-up metadata





Material

Identified tools

Anystyle, CERMINE, EXCITE, GROBID, PDFSSA4MET, Scholarcy, and Science Parse

Data

2,538 bibliographic references: two articles for each one of 27 subject areas in Scimago JR, plus additional two articles having bibliographic references not in a ‘References’/‘Literature’ section

Dimensions

1. Metadata: the number of correctly tagged metadata, independently from content
2. Content: verifying if the text inside a correctly identified metadata is correct
3. Reference: correct if it included the most used metadata (with correct content) for the particular type of the referenced publication (see https://doi.org/10.1007/978-3-031-16802-4_10)

Extraction of bibliographic references, reference metadata, and metadata content of all the tools, including their precision, recall and f-score

Anystyle had the best performance

Lowest f-score was retrieved in the correct identification of references

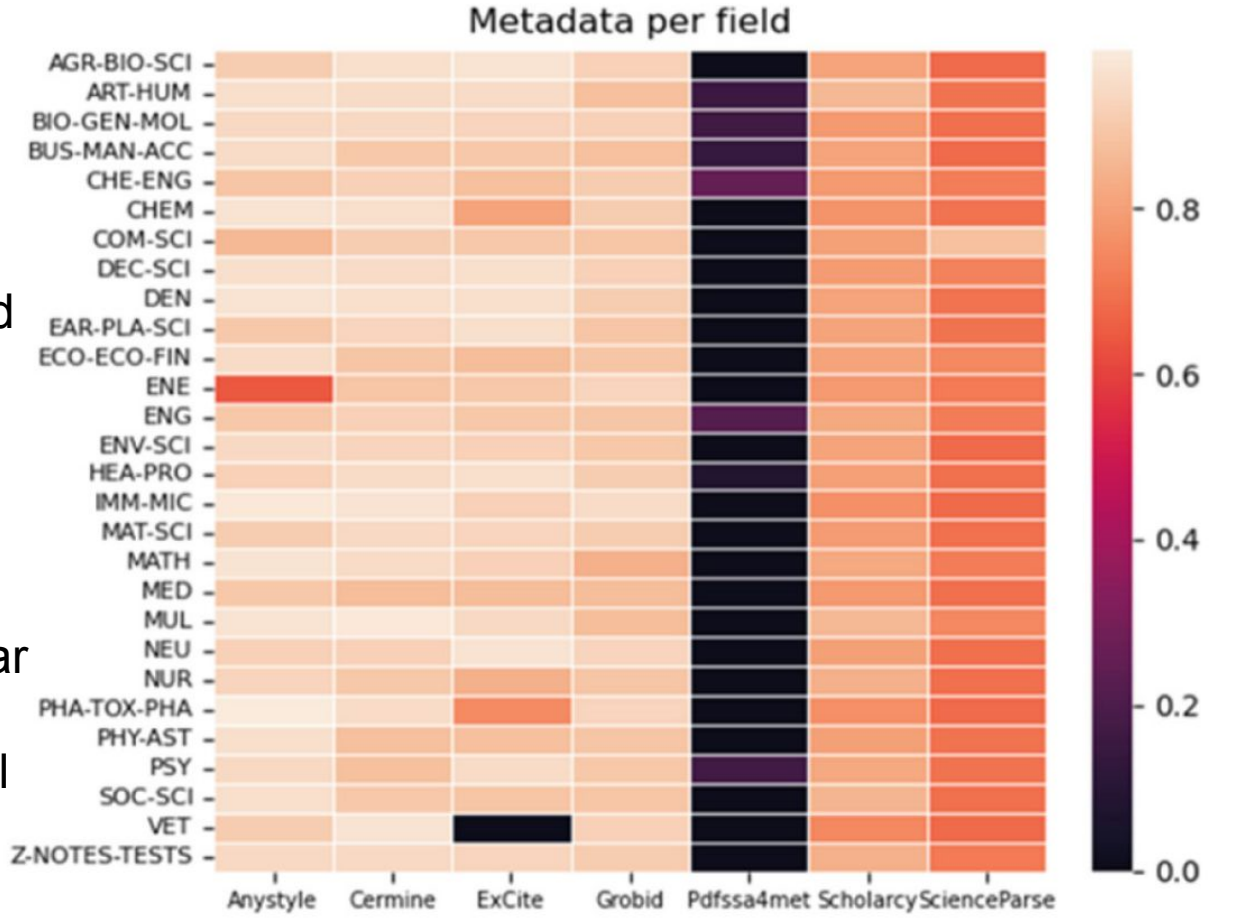
The dimension content showed that, even if the metadata element was correctly identified, the content it contained was prone to parsing errors

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

Metadata

Pdfssa4met was the tool showing the worst performances – identified a few metadata only in seven subject areas and showed a very low precision

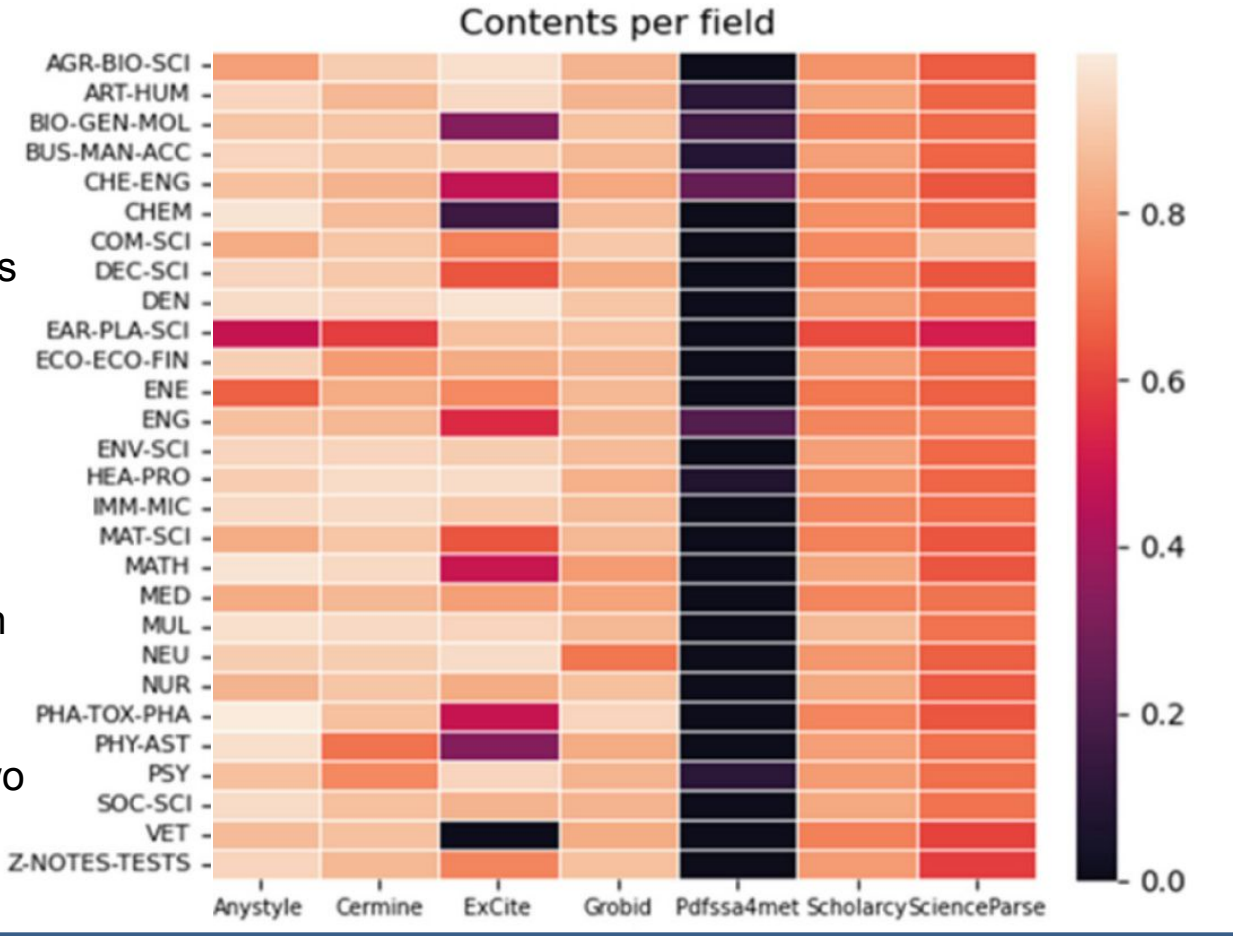
Cermine showed a similar and good f-score for metadata extraction in all disciplines



Content

Some degree of flexibility for string similarity depending on the kind of metadata to assess

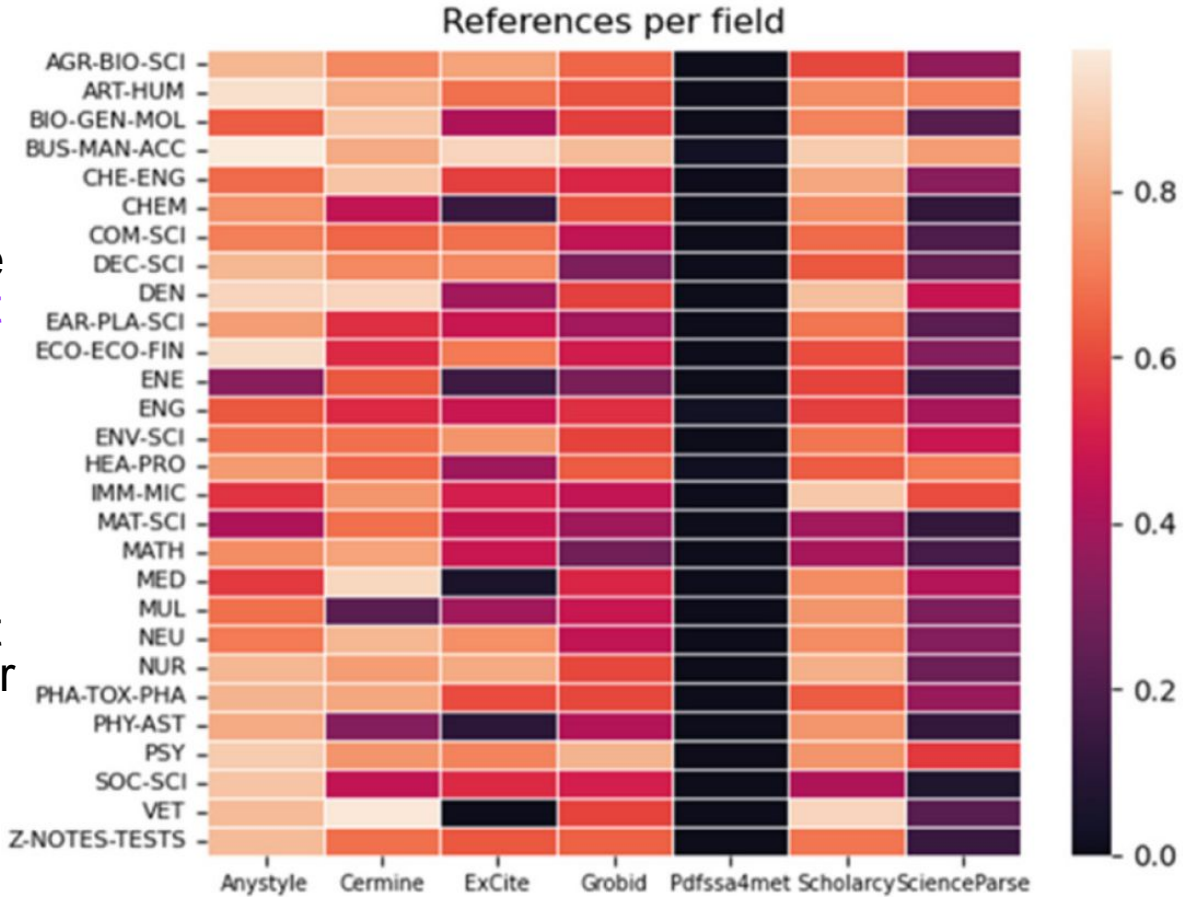
1. strings in bibliographic metadata returned by a tool trimmed
2. non-UTF-8 characters removed
3. comparison via Levenshtein distance
4. different similarity thresholds depending on the type of metadata to compare, i.e. the minimum level of similarity under which two strings could be considered the same



References

Anystyle showed coherent results, with almost all f-values above 0.5 and the highest value registered at 0.97 (BUS-MAN-ACC)

High quality of the identification of references in the set of files which included bibliographic references in a section not labelled as “References” or “Literature” (Z-NOTES-TEST, f-value > 0.85)



Conclusions

We identified only a few tools as functional for the purposes stated in this research

- The lack of maintenance is one of the main reasons why tools that could be considered relevant for our study have been discarded
- Almost all the tools identified are based on machine learning techniques

Anystyle obtained the best f-score in all three dimensions of the analysis, i.e. references, metadata and contents, followed by Cermine – however, the results per subject area showed that, in some cases, Anystyle was outperformed by other tools

Factors that affected the reference extraction by the tools were the citation practice of particular subject areas and the article layout – reference identification was very effective in some subject areas, but other areas (e.g. ENE) showed low performance in all the tools

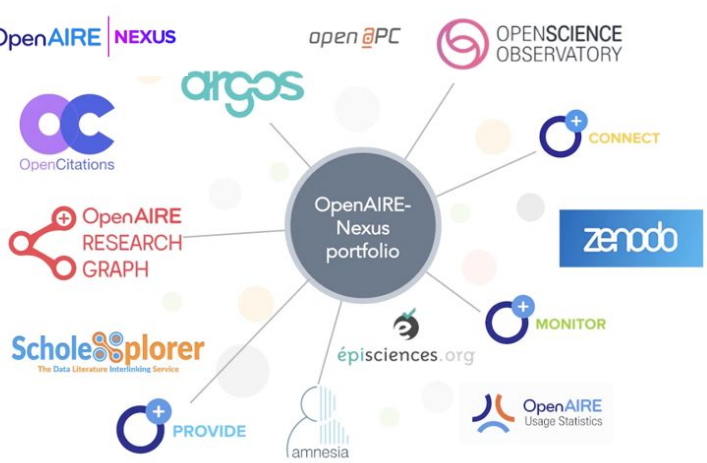
One more thing: context of the work

At the end of 2019, OpenCitations was selected by the Global Sustainability Coalition for Open Science Services (SCOSS) for their second round of crowd-funding support

SCOSS stated that OpenCitations "aligns well with open science goals, is an innovative service, and if successful could be a game changer by challenging established proprietary citation services"

On our blog:
https://opencitations.wordpress.com/2022/01/13/five-reasons-why-2021-has-been-a-great-year-for-opencitations/
https://opencitations.wordpress.com/2022/03/08/opencitations-and-ec-funding-openaire-nexus-and-risis2/
https://opencitations.wordpress.com/2022/05/31/strongtwo-years-of-achievements-within-the-scoss-family-and-its-not-over-yet-strongnbsp/

RISIS
RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES




Thank you
for your attention