The way we cite

Common metadata used across disciplines for defining bibliographic references

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

Erika Alves dos Santos, Silvio Peroni*,Marcos Luiz Mucheroni

*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

Source: https://www.autodeskresearch.com/publications/citeology

Citations are fundamental tools to track how science evolves over time

They link scientific thinking forming a complex chain of documents highlighting of research trends

Bibliographic references enables the creation of such a conceptual link, and carry an important function: providing enough metadata to facilitate an agent to identify the cited works

Citation: a conceptual directional link from a citing entity to a cited entity, for the purpose of acknowledging or ascribing credit for the contribution made by the author(s) of the cited entity.

The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles

Research article | Legal Issues | Science Policy | Data Science

Published February 13, 2018

Author and article information

Abstract

REFERENCES

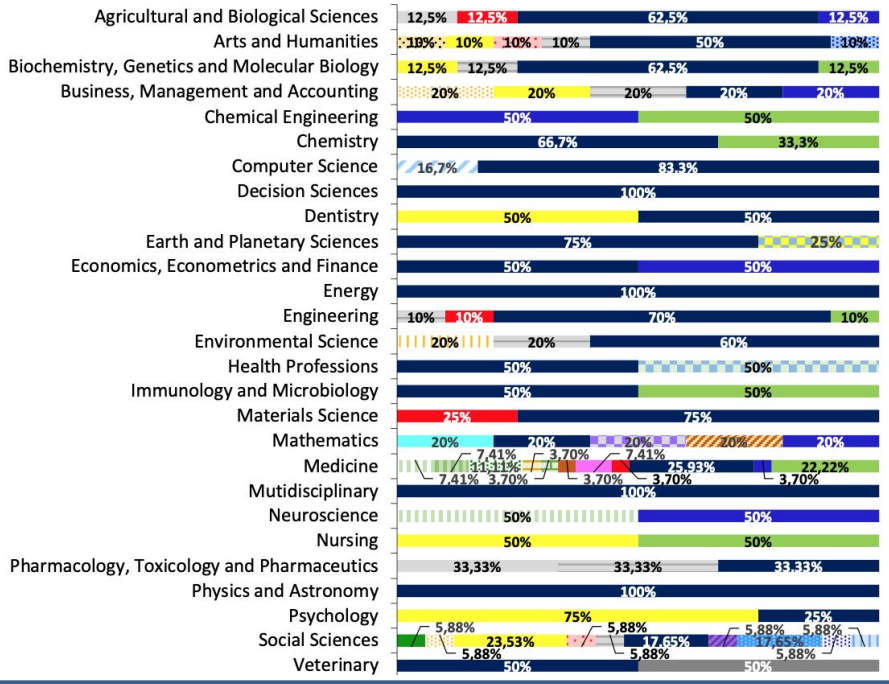
Björk BC, Laakso M, Welling P, Paetau P. 2014. Anatomy of green open access. *Journal of the Association for Information Science and Technology* **65(2)**:237–250.

Citation behaviour: theory vs practice

Theory
In the past, a huge number of citation style manuals (APA, Chicago, Harvard, etc.) have been released to provide standardised approaches to the definition of bibliographic references – and, in particular, their metadata

Practice
The citation practices observed in the published literature are very noisy, confusing, and not standardised at all – e.g. several journals often avoid adopting standardised citation style manuals and define their **own (yet another) citation style**

dos Santos, Peroni, Mucheroni (2022). An analysis of citing and referencing habits across all scholarly disciplines: approaches and trends in bibliographic metadata errors. arXiv. https://doi.org/10.48550/arXiv.2202.08469



Analysing current practices

Three research questions:

RQ1. Which entities are cited by articles published in journals of different disciplines?

RQ2. What is the standard metadata set used across such disciplines for describing cited works within bibliographic references?

RQ3. Is there any mechanism in place (i.e. hypertextual links) to facilitate the algorithmic recognition of where a bibliographic reference is cited in the text?

Methodology (data manually extracted and curated)

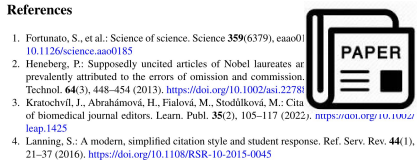
SJR
Scimago Journal & Country Rank

Selected the most cited journals in each of the 27 subject areas listed in SCImago in the 2015–2017 triennium according to the SCImago total cites ranking

Grouped in 5 macro categories (Health Sci., SSH, Life Sci., Physical Sci., Multidisciplinary)

Each journal in the sample was represented by five articles (in PDF format) published in the most recent issue published between in October 2019

729 articles (172 Health Sci., 191 SSH, 114 Life Sci., 232 Physical Sci., 20 Multidisciplinary)



We extracted all the bibliographic references in the bibliographic references lists of the selected journal articles

34,140 bibliographic references

Metadata

We detected the types of the cited works in each discipline and the structure of bibliographic references for each type of cited work

36 different types of cited publications, 64 different kinds of metadata

Santos, Peroni, Mucheroni (2020). Workflow for retrieving all the data of the analysis introduced in the article “Citing and referencing habits in Medicine and Social Sciences journals in 2019”. https://doi.org/10.17504/protocols.io.bbifikb

Cited types (RQ1): rationale

We considered all the 34,140 bibliographic references composing our sample, that we used to identify the following different kinds of publications:

articles, books and related chapters, manuscripts, technical reports and related chapters, webpages, proceeding papers, conference papers, grey literature, data sheets, forthcoming chapters, forthcoming articles, unpublished material, standards, working papers and preprints, e-books and related chapters, newspapers, online databases, web videos, patents, software, manuals/guides/toolkits, personal communications, book series, memorandum, governmental official publications, legislation, informative materials, audio records, motion pictures, speeches, photographs, slide presentation, podcasts, engravings, lithography, and television shows

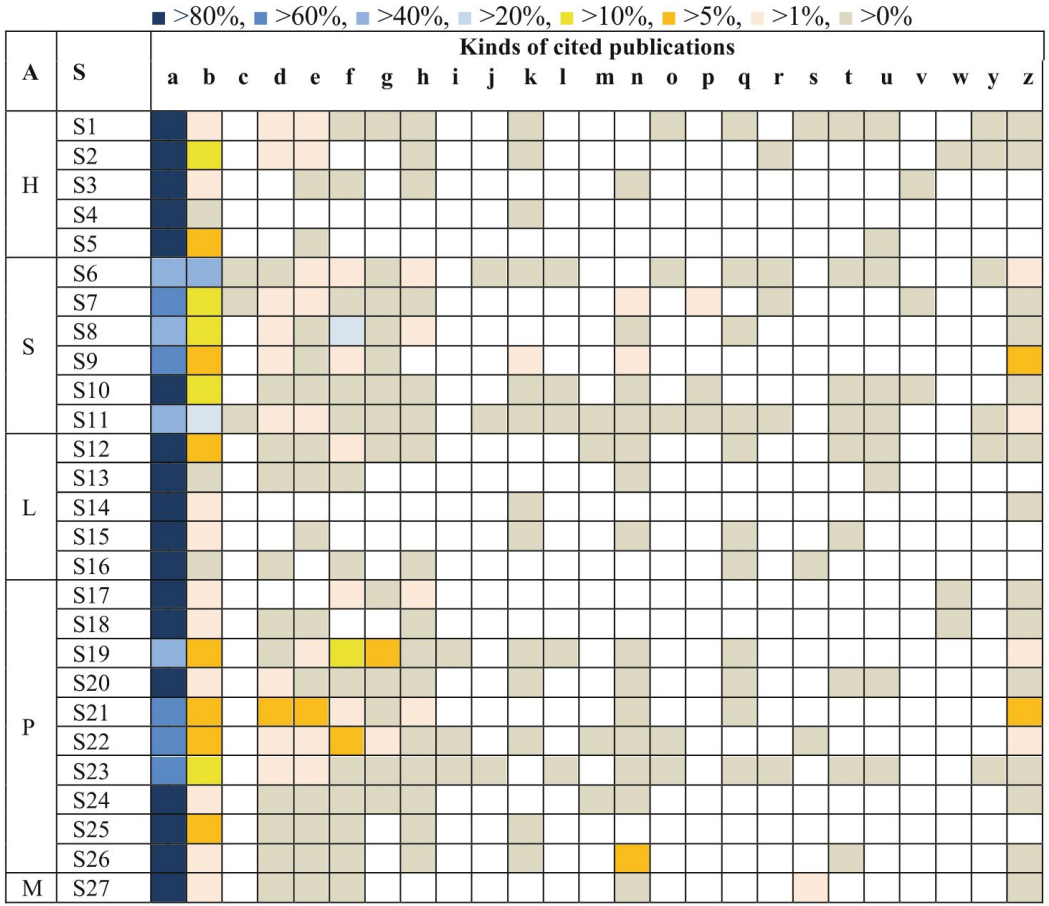
Santos, E.A.d., Peroni, S., Mucheroni, M.L.: Raw and aggregated data for the study introduced in the paper “The way we cite: common metadata used across disciplines for defining bibliographic references” (2022). https://doi.org/10.5281/zenodo.6586859

Cited Types (RQ1): results

Articles [a], books (and their chapters) [b], and proceeding papers [f] were the first, second and third most cited types of publications across all the subject areas

Grey literature [h] is the eighth most cited type of work across all subject areas (0.47% of total bibliographic references), the third most cited type of publication in arts and humanities articles (S6) and the fourth most cited type in chemical engineering (S17), decision sciences (S8) and mathematics (S25)

Some subject areas (e.g. social sciences – S11) tend to cite a greater variety of types of publications while others (e.g. dentistry – S4) only a few types



Metadata (RQ2): rationale

We decided to select the seven most cited types of publications in each subject area – all the types of publications were considered except manuscripts (c), forthcoming chapters (j), web videos (r), other kinds (y) and unidentified types of publications (z)

33,786 bibliographic references were individually analysed to identify their descriptive elements (i.e. metadata)

We marked all the elements specified in at least one bibliographic reference of at least 50% of the articles composing each subject category

We computed the most used descriptive elements for each type of publication mentioned above by considering each macro area's most used descriptive elements

A descriptive element was selected if it was one of the most used in all the macro areas

Metadata (RQ2, part 1): results

The title (11) is one of the most used metadata across all the macro areas, but in 27% of articles from Physical Sciences, bibliographic references pointing to web pages did not provide the title of the cited work

The DOI (61) is not included in the most used metadata (row A) in the bibliographic references referring to articles, as the ISBN (39) is not part of the most used metadata in the bibliographic references referring to books and book chapters

Articles		Books		Book chapters	
H	4,11,17,22,33,50	H	11,17,18,30,32,33	H	1,11,14,17,30,32,33,48,64
S	5,11,17,19,22,33,50	S	11,17,30,32,33	S	1,11,14,17,30,32,33,48,64
L	4,11,17,22,33,36,50	L	11,17,30,32,33	L	1,11,14,17,30,32,33,48,64
P	4,11,17,22,33,50	P	11,17,30,32,33	P	1,11,14,17,30,32,33,48,64
M	4,11,17,22,33,36 50	M	11,17,32,33	M	1,11,14,17,30,32,33,38,48,64
A	**11,17,22,33,50**	**A**	**11,17,32,33**	**A**	**1,11,14,17,30,32,33,48,64**

Technical reports		Webpages		Proceeding papers	
H	4,11,17,22,33,50	H	11,17,26,60	H	3,11,17,33,48,64
S	5,11,17,19,22,33,50	S	11,17,33,60	S	8,11,17,30,32,33,48,64
L	4,11,17,22,33,36,50	L	11,17,33,60	L	8,11,17,32,33,48,64
P	4,11,17,22,33,50	P	11,17,33,60	P	8,11,17,32,33,48,64
M	4,11,17,22,33,36,50	M	11,17,60	M	3,8,11,17,32,33,48,64
A	**11,17,22,33,50**	**A**	**11,17,60**	**A**	**11,17,33,48,64**

Conference papers		Grey literature		Data sheets		Technical rep. chapters	
H	3,11,17,33	H	11,17,30,32,33,46	H	No citations	H	1,11,14,22,30,32,33,60
S	3,11,17,25,33	S	11,17,32,33,46	S	No citations	S	1,11,14,30,32,33
L	3,11,17,25,33	L	11,17,30,32,33,46	L	No citations	L	No citations
P	3,11,17,25,33	P	11,17,30,32,33,46	P	11,32,33	P	No citations
M	No citations	M	No citations	M	No citations	M	No citations
A	**3,11,17,33**	**A**	**11,17,32,33,46**	**A**	**11,32,33**	**A**	**1,11,14,30,32,33**

Metadata (RQ2, part 2): results

Bibliographic references should provide (at least) the necessary metadata for the proper identification of the referred publications

Despite the existence of thousands of reference styles and standards to guide the use and interpretation of bibliographic metadata uniformly, the same type of publication may have different descriptions in different disciplines

Forthcoming articles	
H	4,11,17,33,58,61
S	5,11,17,58
L	4,11,17,22,29,33,58,60
P	4,11,17,33,58
M	No citations
A	**11,17,58**

Unpublished	
H	No citations
S	11,17,33,57
L	No citations
P	11,17,32,33,57
M	No citations
A	**11,17,33,57**

Standards	
H	No citations
S	11,17,33
L	11,17,30,33
P	11,17,18,33,51
M	No citations
A	**11,17,33**

Working papers	
H	11,17,26,30,32,33,60
S	11,17,33,45,60
L	11,17,26,33,61
P	11,17,33,60
M	11,17,32,33,60
A	**11,17,33**

E-books	
H	11,17,30,32,33
S	11,17,30,32,33
L	11,17,26,30,32,33,60
P	11,17,18,26,33,39,60,61
M	No citations
A	**11,17,33**

Newspapers	
H	No citations
S	7,11,17,28,33,60
L	No citations
P	No citations
M	No citations
A	**7,11,17,28,33,60**

Online databases	
H	11,17,21,26,33,60
S	11,17,32,33,60
L	11,17,21,33
P	11,17,21,32,33,46,60,61
M	No citations
A	**11,17,33**

E-books chapters	
H	1,11,14,17,30,32,33,48
S	1,11,17,30,32,33,64
L	No citations
P	1,11,14,17,26,33,60,64
M	No citations
A	**1,11,17,33**

Patents	
H	11,17,33,41
S	No citations
L	11,17,33,41,46
P	11,17,30,33,41,48
M	11,17,33,41,60
A	**11,17,33,41**

Software	
H	11,17,30,32,33
S	11,17,30,32,33,46
L	11,17,21,26,30,32,33,60
P	11,17,21,33,60
M	No citations
A	**11,17,33**

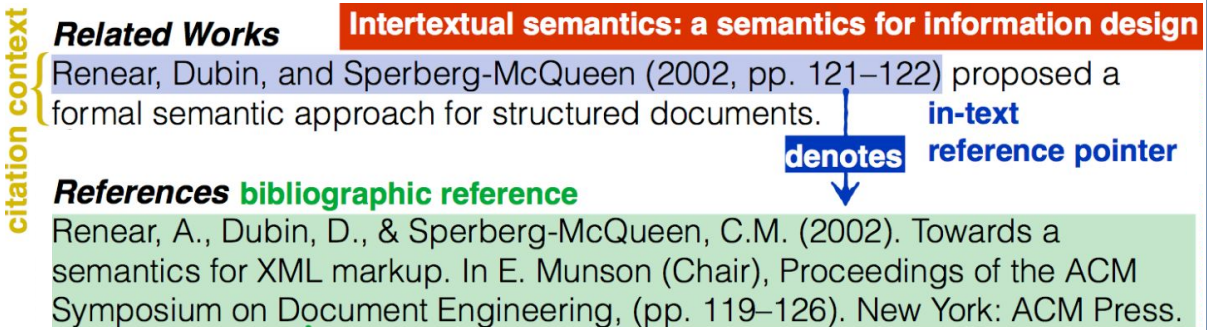
Manual/guides/toolkits	
H	11,17,30,32,33,60
S	11,17,30,32,33
L	11,17,32,33
P	11,17,21,32,33
M	No citations
A	**11,17,32,33**

Personal communications	
H	11,17,30,32,33,60
S	11,17,28,33,46,60
L	No citations
P	No citations
M	No citations
A	**11,17 33,60**

Book series	
H	1,14,19,22,32,33,34,47,49,61
S	No citations
L	No citations
P	1,14,19,22,32,33,34,49,61
M	No citations
A	**1,14,19,22,32,33,34,49,61**

Linked in-text reference pointers (RQ3): rationale

Considering all the in-text reference pointers – e.g. “(Doe et al., 2022)” and “[3]” – denoting all the bibliographic references in our sample



Goal: how many of them are accompanied by a link pointing to the related bibliographic reference they denote

Such links are helpful tools to formalise the connections between the text of the citing article and the correspondent cited works referenced by the bibliographic references

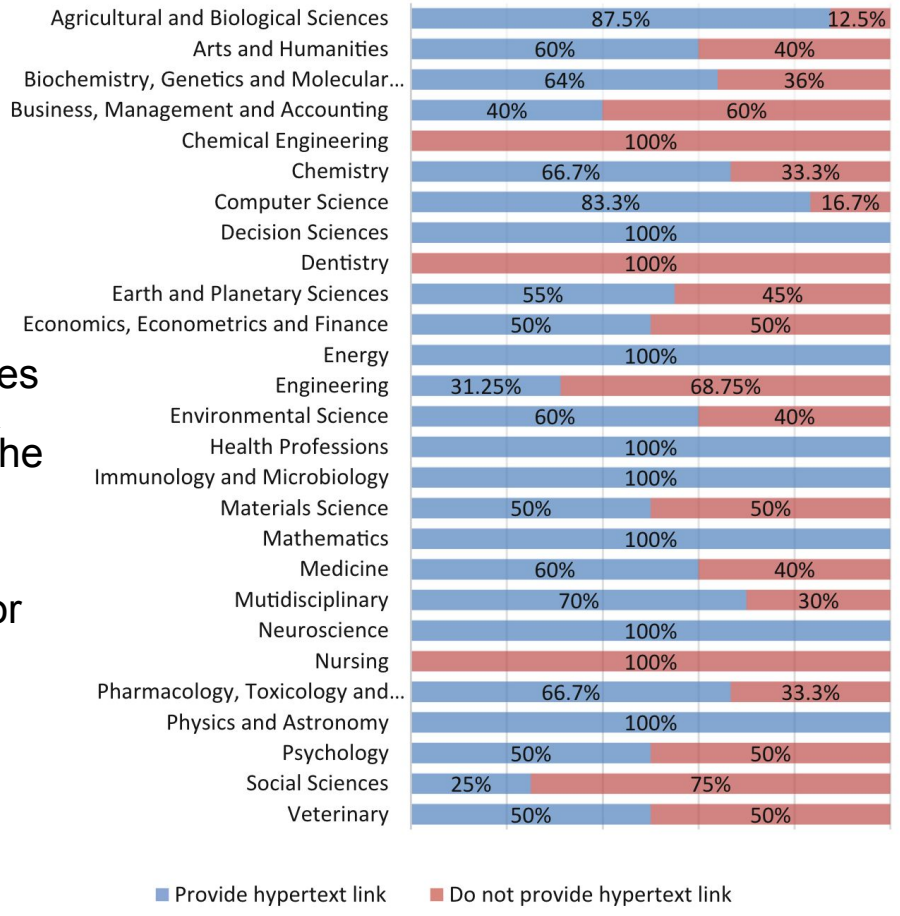
Linked in-text ref. point. (RQ3): results

49% of the articles provide such links

Having such such mechanisms in place simplifies the development of computational tools to track where cited works are referred to in the text of the citing articles, thus facilitate the computational recognition of:

- citation sentences
- citation functions , i.e. the reason an author cites a cited work

If no links are available, natural language processing tools and other techniques can be used – but it is more complex due to the heterogeneity of the formats for bibliographic references and in-text reference pointers



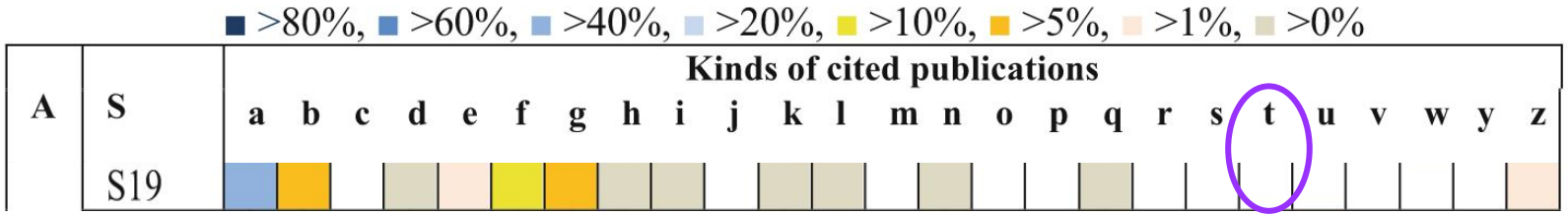
Conclusions

Take away messages:

- 36 different types of cited works, but there exists different citing behaviours in scientific articles that varied from subject area to subject area
- Description of different types of publications may demand different types of metadata, which do not necessarily play the same role in the identification of the cited work

A question for future works (I personally care of):

Why software (t) was not listed among the most cited type of works in Computer Science (S19) while being one of the main topics discussed in several areas of Computer Science research?



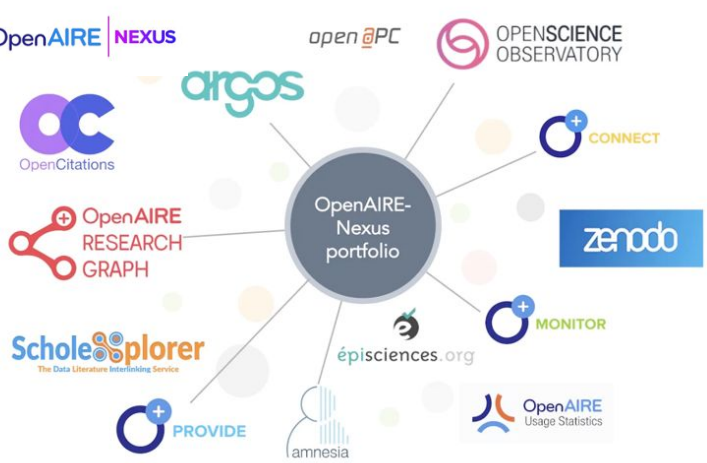
One more thing: context of the work

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