OpenCitations

Open citations in Informatics

European Computer Science Summit 2021

Silvio Peroni

Director of OpenCitations, Director of the Research Centre for Open Scholarly Metadata

Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy

silvio.peroni@unibo.it – @essepuntato – contact@opencitations.net – @opencitations

RESEARCH CENTRE
FOR OPEN SCHOLARLY METADATA

OpenCitations

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

What is an *open* citation

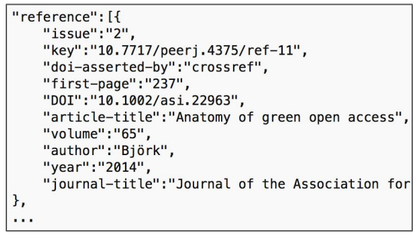
Citation: a conceptual directional link from a citing entity to a cited entity, for the purpose of acknowledging or ascribing credit for the contribution made by the author(s) of the cited entity.

The citation data related to a particular citation must include:

- the *representation* of such a conceptual directional link
- the *basic metadata* of the citing entity and the cited entity, i.e. sufficient information to create or retrieve textual bibliographic references

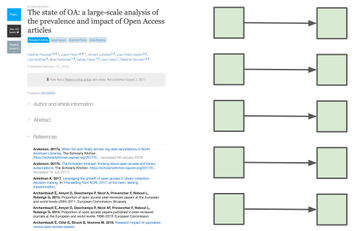
A citation is open when its data are compliant with the following characteristics:

structured



```
"reference":[{
    "issue":"2",
    "key":"10.7717/peerj.4375/ref-11",
    "doi-asserted-by":"crossref",
    "first-page":"237",
    "DOI":"10.1002/asi.22963",
    "article-title":"Anatomy of green open access",
    "volume":"65",
    "author":"Björk",
    "year":"2014",
    "journal-title":"Journal of the Association for
},
...
```

separate



open

without restrictions to maximise reuse

identifiable

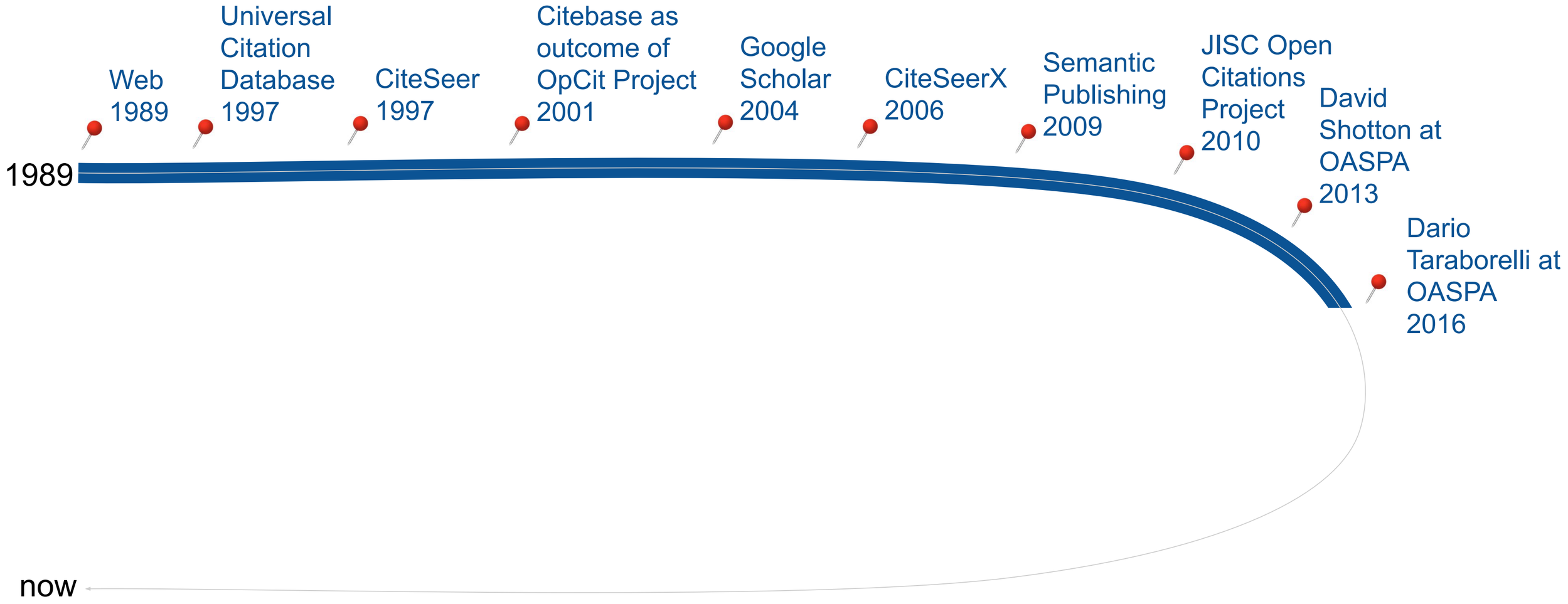
PID

available

http:// + PID = metadata

S. Peroni & D. Shotton (2018).
Open Citation: Definition. https://doi.org/10.6084/m9.figshare.6683855

History of open citations, in brief (1/3)



History of open citations, in brief (2/3)

1989

Web
1989

Universal
Citation
Database
1997

CiteSeer
1997

Citebase as
outcome of
OpCit Project
2001

Google
Scholar
2004

CiteSeerX
2006

Semantic
Publishing
2009

JISC Open
Citations
Project
2010

David
Shotton at
OASPA
2013

Dario
Taraborelli at
OASPA
2016

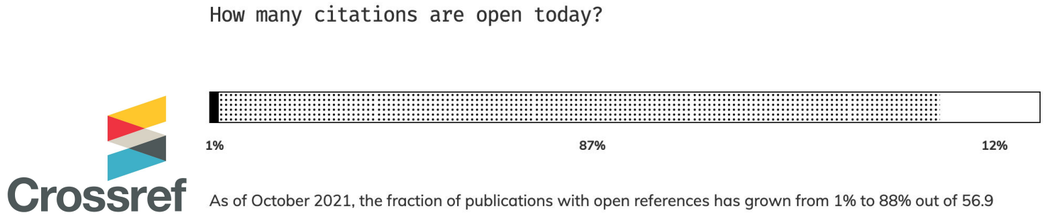
OpenCitations
WikiCite
2016

Initiative for
Open Citations
2017

WikiCite
Workshop
2017

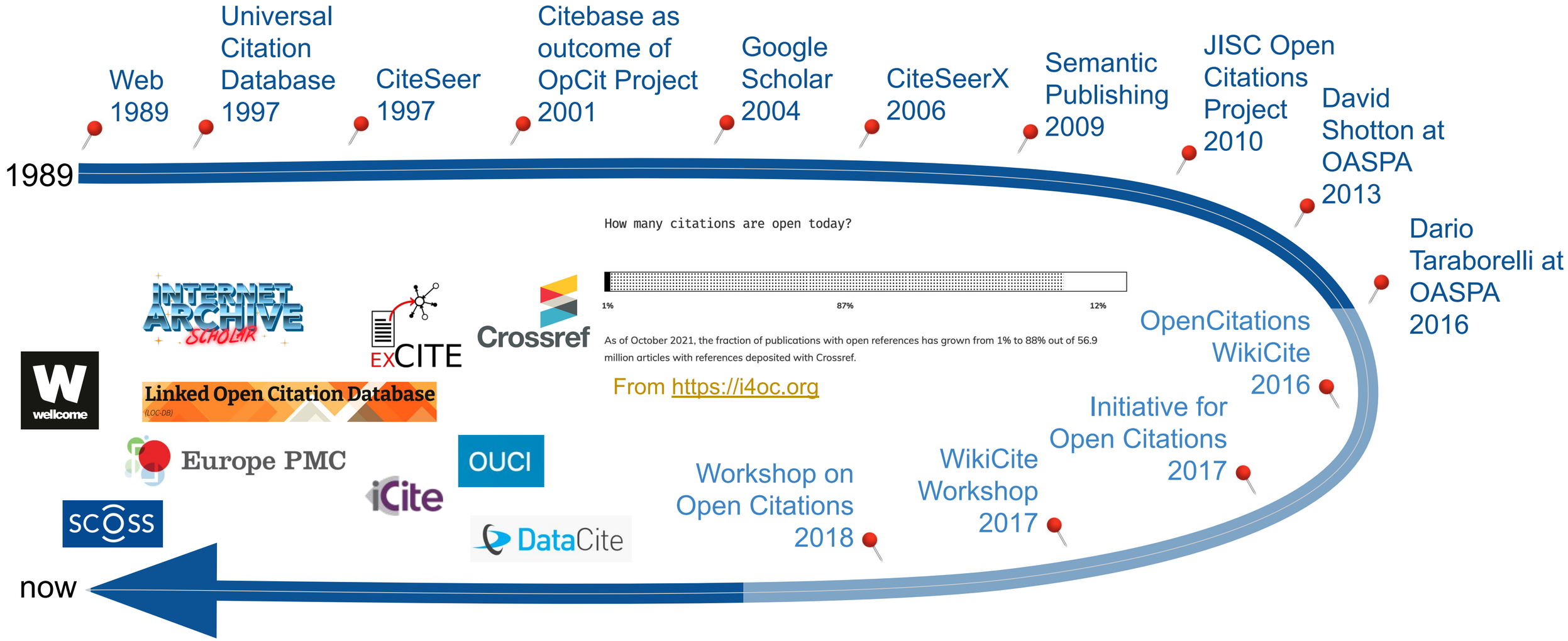
Workshop on
Open Citations
2018

now



From https://i4oc.org

History of open citations, in brief (3/3)



OpenCitations: basic information

OpenCitations (http://opencitations.net) is an independent not-for-profit infrastructure organization

- dedicated to open scholarship and the publication of open bibliographic and citation data by the use of Semantic Web (Linked Open Data) technologies
- engaged in advocacy for open citations and open bibliographic metadata, as a founding member of both the Initiative for Open Citations (I4OC) and the Initiative for Open Abstracts (I4OA)

It provides:

- a data model: the OpenCitations Data Model (based on the SPAR Ontologies)
- bibliographic and citation databases (CC0): OpenCitations Corpus, OpenCitations Indexes, Open Biomedical Citations in Context Corpus
- software: in our GitHub repository, released with open source licenses
- online services: REST APIs, SPARQL endpoints, dumps, and search and query interfaces

OpenCitations: values

Fairness: avoid that institutions and independent scholars having to pay tens of thousands of euros annually (that most of them cannot afford!) for relevant services and data, e.g. commercial access to their own scholarly data

Reuse: licenses to foster maximum discoverability and reuse, so users can re-publish and use, for any purpose, data and software that we provide

Transparency: providing open and freely available knowledge of relevance for the scholarly ecosystem (e.g. for dissemination activities and within research evaluation exercises) to foster transparent and reproducible processes

Governance: crucial to involve the stakeholder community in the governance, to build more confidence that the organisation will take decisions driven by community consensus and consideration of different interests

OpenCitations Indexes

OpenCitations Indexes (http://opencitations.net/index), including COCI (launched in July 2018), the OpenCitations Index of Crossref open DOI-to-DOI citations, currently contains more than

1.18 billion citations between 69 million bibliographic resources

COCI data and services	URL
REST API *results in CSV and JSON*	https://opencitations.net/index/coci/api/v1
SPARQL endpoint *results in several formats*	https://opencitations.net/index/sparql
Dumps *all data in CSV, JSON, RDF*	http://opencitations.net/download#coci

Supporting Open Science Infrastructures

The Principles of Open Scholarly Infrastructure

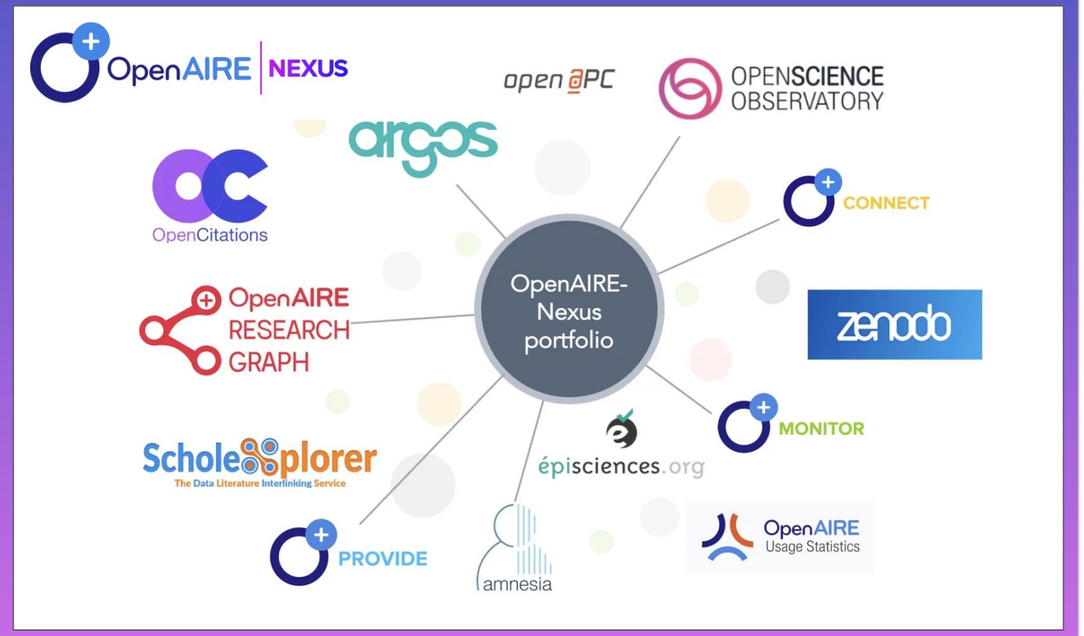
Governance	● ● ● ● ● ● ●	Improve stakeholder involvement in governance and plan for winding-down
Sustainability	● ● ● ● ●	Not enough incomes to generate surplus and contingency fund to support 12 month operations
Insurance	● ● ● ●	Nothing to declare :-)

DIRECTORY OF OPEN ACCESS BOOKS
OPEN ACCESS PUBLISHING FOR EUROPEAN NETWORKS
PUBLIC KNOWLEDGE PROJECT
OPENCITATIONS

SCOSS LAUNCHES SECOND FUNDING CYCLE

Read about these essential services, their funding goals and how your institution can help support them at www.scoss.org.

SCOSS



Some experiments

Goals:

- How many citations involving a Computer Science publications are included in OpenCitations?
- What are the publishers which are contributing the most?
- What are the publishers receiving more citations by the others?

Material:

- Citations included in COCI September 2021 Dump (https://doi.org/10.6084/m9.figshare.6741422.v11)
- Bibliographic metadata of articles in DBLP October Dump (https://dblp.org/xml/release/dblp-2021-10-01.xml.gz)
- Publisher information via Crossref API (https://api.crossref.org) and DataCite API (https://api.datacite.org)

Data gathered

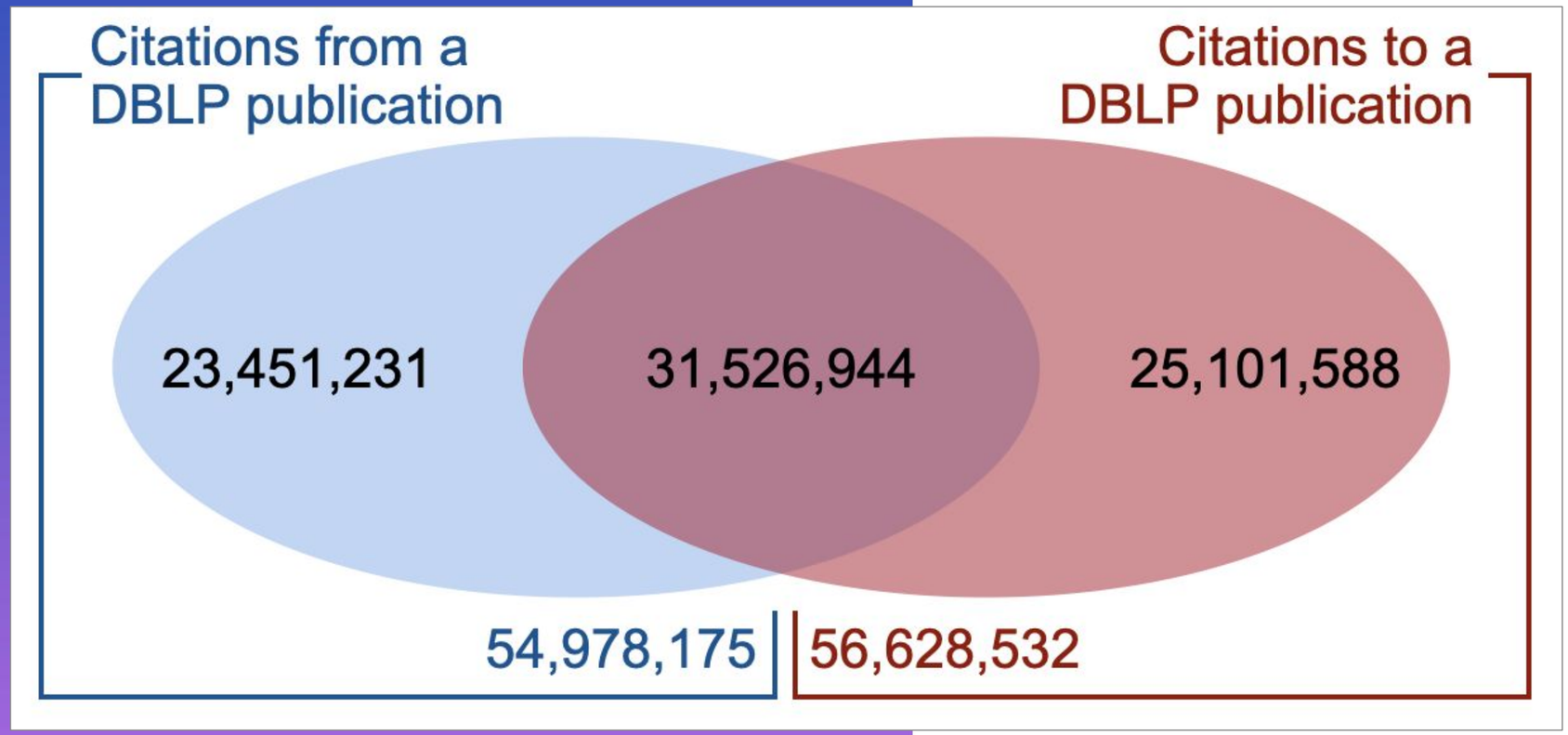
DBLP used for certifying Computer Science publications (it is an approximation, of course):

- 4,637,865 entities in DBLP (only journal articles, conference proceedings, books and book chapters) with a DOI

COCI used to retrieve all (DOI-to-DOI) citations:

- 80,079,763 citations in COCI that involve a DBLP entity in the previous set
- ~11.8 citations per DBLP citing entity (COCI overall average: ~23.2 citations per citing entity)

All data are available at https://doi.org/10.5281/zenodo.5595388
Software used to gather the data at https://github.com/essepuntato/ecss-2021

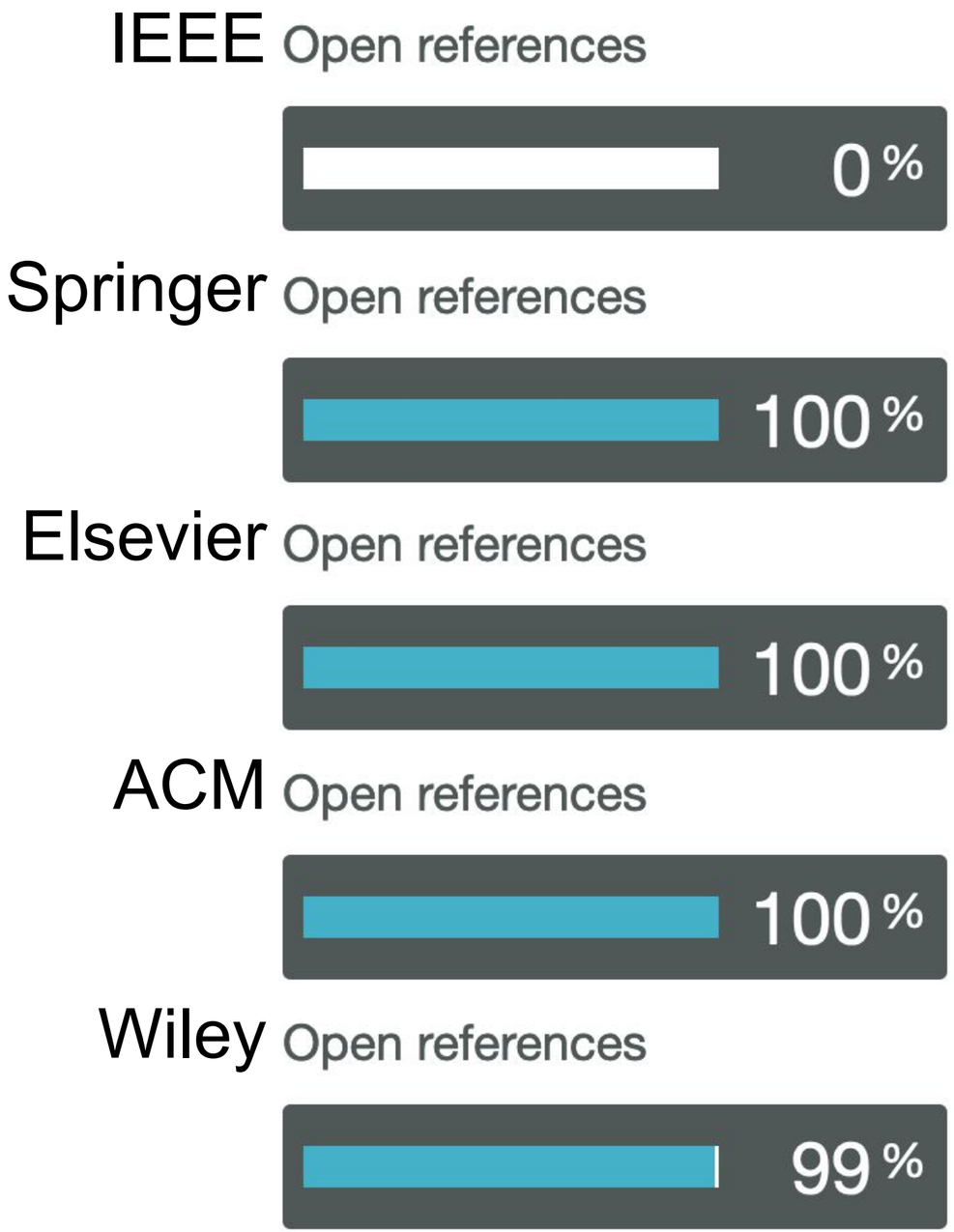


31% of citations come from publications not included in DBLP

(from other disciplines?)

Future analysis:
Who cites Informatics?

Publisher	Entities	Outgoing	Incoming
IEEE	1,730,485	18,930,055	21,582,093
Springer	1,012,534	18,482,132	11,179,566
Elsevier	574,860	15,536,207	17,019,716
ACM	433,188	3,695,255	6,050,342
Wiley	89,662	3,350,183	3,357,065



source: https://www.crossref.org/members/prep/

Italian NSQ: citation coverage

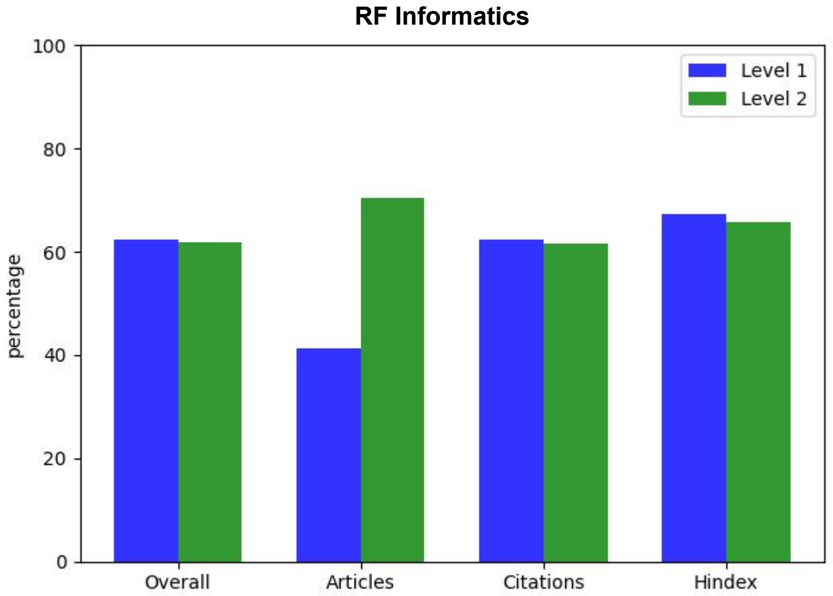
Italian National Scientific Qualification (NSQ) is a nation-wide research assessment exercise which establishes whether a scholar can apply to professorial academic positions Full Professor (Level 1) and Associate Professor (Level 2)

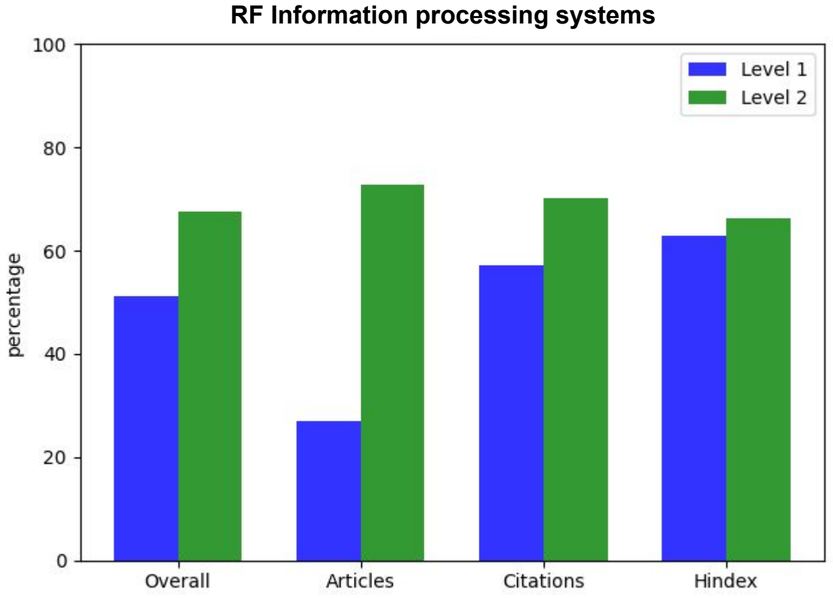
Applications are organized according to a governmentally-defined taxonomy of 190 Recruitment Fields (RFs), each assigned to one of the following categories, citation-based (CD) and non-citation-based (ND), depending on the use of citations for the evaluation

Web of Science and Scopus are used to gather bibliographic metadata and citations

The coverage of open citations (from COCI December 2020 Dump) not sufficient to replace proprietary databases

Future analysis: repeat the experiment with new COCI dump, that also include Elsevier citations (missing in the previous experiment)





F. Bologna, A. Di Iorio, S. Peroni & F. Poggi (2021). Can we assess research using open scientific knowledge graphs? A case study within the Italian National Scientific Qualification. https://arxiv.org/abs/2105.08599

Italian NSQ: article coverage

The 190 RFs are organised into 14 Scientific Areas (SAs)

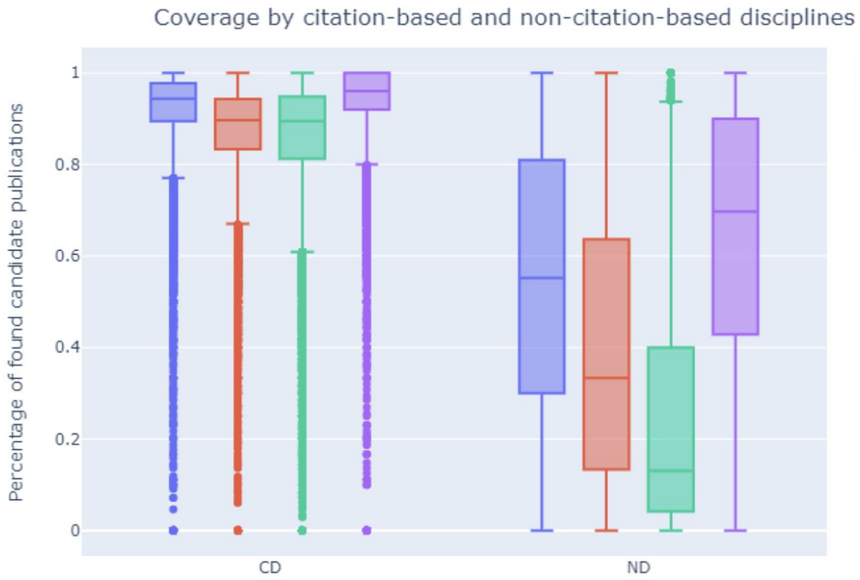
When preparing their applications, candidates can select only publications in Scopus and Web of Science

We wanted to check if we have similar coverage if we use only open metadata retrieved from:

- Microsoft Academic Graph (MAG)
- OpenAIRE (OA)
- Crossref (CR)

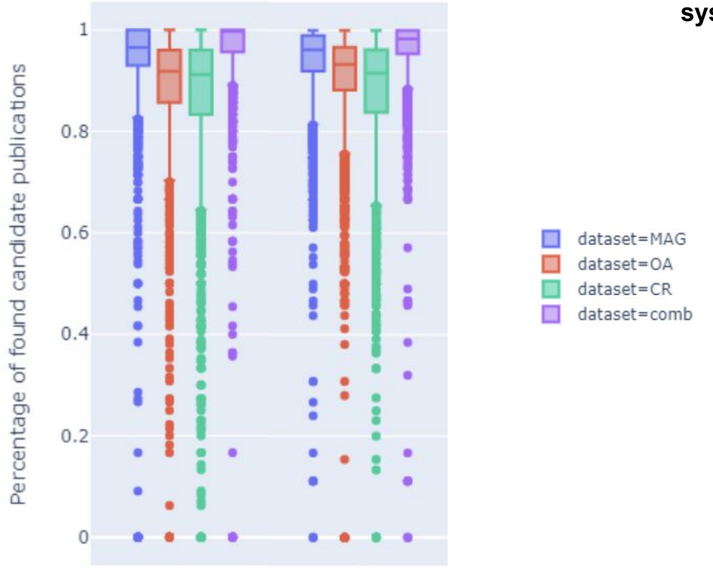
CD RFs show better coverage than ND RFs

Very good results for NSQ applications in the SAs including RF Informatics and Information processing systems (very close to 100% coverage compared with Scopus and Web of Science in both the SAs)



SA Mathematics and Informatics (includes RF Informatics)

SA Industrial and Information Engineering (includes RF Information processing systems)



F. Bologna, A. Di Iorio, S. Peroni & F. Poggi (2021). Open bibliographic data and the Italian National Scientific Qualification: measuring coverage of academic fields. https://arxiv.org/abs/2110.02111

- Open citations are available without charge and under open licences
- Goal: data with scope, depth, accuracy and provenance, as a disruptive alternative to traditional proprietary citation indexes

More than 1.18 billion open citations available (even more, if you consider all the existing providers in addition to OpenCitations)

Tools (e.g. API) supporting easy integration with systems

Dumps available for huge analysis and usage

Still missing some major publishers

DOI-to-DOI citations do not include several citations, e.g. CEUR Workshop Proceedings citations are all missing in OpenCitations

More (coordinated?) effort needed to have a coverage comparable to proprietary services

Thanks for your attention