Workshop on Open Citations: Opportunities and Ongoing Developments, September 2
ISSI2019, September 2-5, 2019, Sapienza University, Rome, Italy

OpenCitations

An infrastructure for open citation data

Ivan Heibi
Digital Humanities Advanced Research Centre (DHARC),
Department of Classical Philology and Italian Studies,
University of Bologna, Bologna (Italy)
ivan.heibi2@unibo.it

OpenCitations

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

About me

...-2017

Bachelor and Master degree in Computer Science at the **University of Bologna, Italy**

2017-18

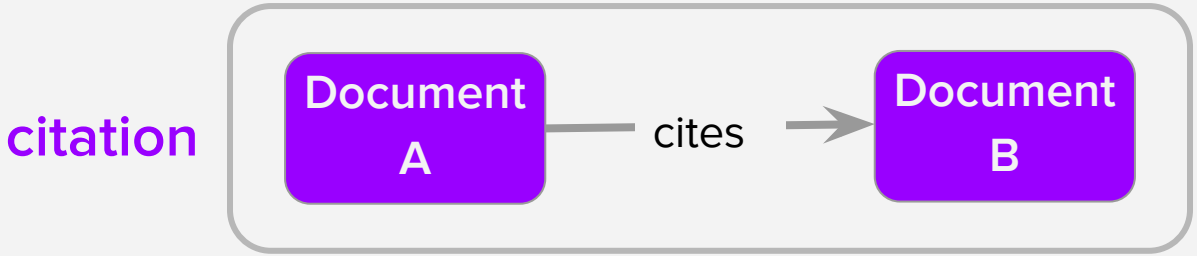
Research fellow under the supervision of **Silvio Peroni** on the "**OpenCitations** Enhancement Project", funded by the Alfred P. Sloan Foundation

2018-...

Ph.D. student at the Department of Classical Philology and Italian Studies, University of Bologna, Bologna, Italy;
The application of Semantic Publishing technologies in the Science of Science research domain

A Citation

- A **bibliographic citation** is a conceptual directional link from a **citing entity** to a **cited entity.** N.B. This definition is **entity-independent.**



- 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 bibliographic citation is an **open citation** when the data needed to define the citation are: **structured, separate, open, identifiable, available**

Open citations: characteristics

Identifiable

Available
E.g. via HTTP


```
"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
},
...
```


Structured
(JSON; machine readable)

https://api.crossref.org/works/10.7717/peerj.4375

Separate
(e.g. via REST calls)

"Estimation of WOS costs is about $100,000 per year for large organizations [...] the cost of Scopus database is about 85-95% of the cost of WOS for the same organizations"
https://doi.org/10.5539/ass.v9n5p18

Closed

Open

"No claims of ownership to individual items of bibliographic metadata"
https://api.crossref.org

OpenCitations

OpenCitations (http://opencitations.net) is a **scholarly infrastructure organization**

It works on:

- the **publication of open bibliographic and citation data** by the use of **Semantic Web technologies,** and RDF for its description
- **advocacy for open citations**

It provides:

- **data models**: the OpenCitations Data Model (based on SPAR Ontologies)
- **datasets (in CC0)**: OpenCitations Corpus, COCI, CROCI
- **software**: GitHub repository released with open source licenses
- **online services**: dumps, REST APIs, SPARQL endpoints, and interfaces

Peroni, Shotton (2019). OpenCitations. arXiv. https://arxiv.org/abs/1906.11964

Semantic Web technologies essentials

- What is it ?
  An extension of the World Wide Web -> **the web of data** that can be processed by machines

- How is that possible ?
  Resources on the web are described using the **RDF data model: a Triple (subject, predicate, object)**

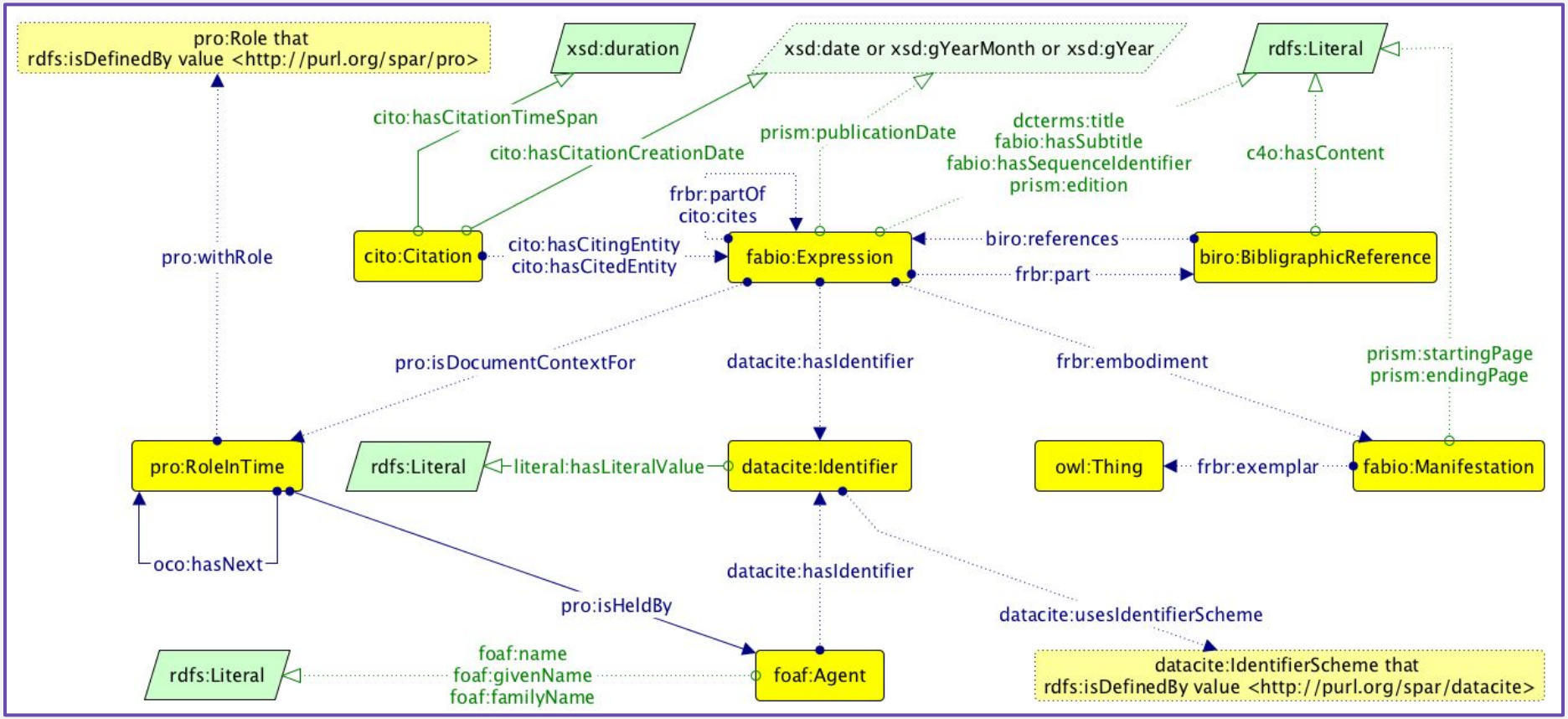
<https://w3id.org/oc/corpus/br/1> <http://purl.org/spar/cito/cites> <https://w3id.org/oc/corpus/br/79> .

RDF example from OpenCitations

- Tim Berners-Lee rules for **Linked open data:**
  - Use **HTTP URIs to name all the resources**, so that they could be **looked up (interpreted)**.
  - Provide useful information about what a name identifies when it's looked up: **metadata**
  - **Refer to other things using their HTTP URI**-based names when publishing data on the Web.

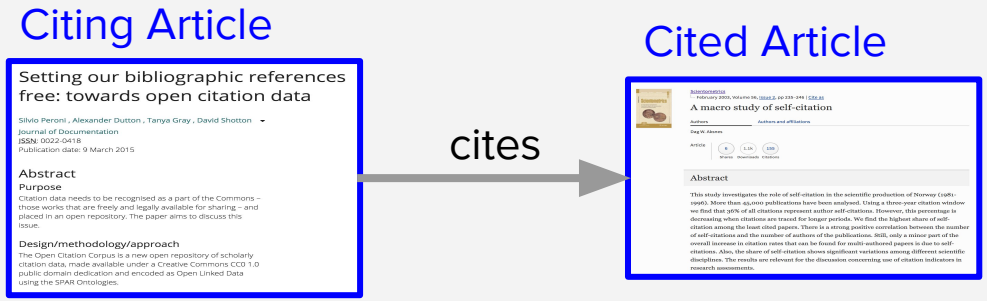
Tim Berners-Lee on the next Web, https://www.ted.com/talks/tim_berners_lee_on_the_next_web

OpenCitations Data Model

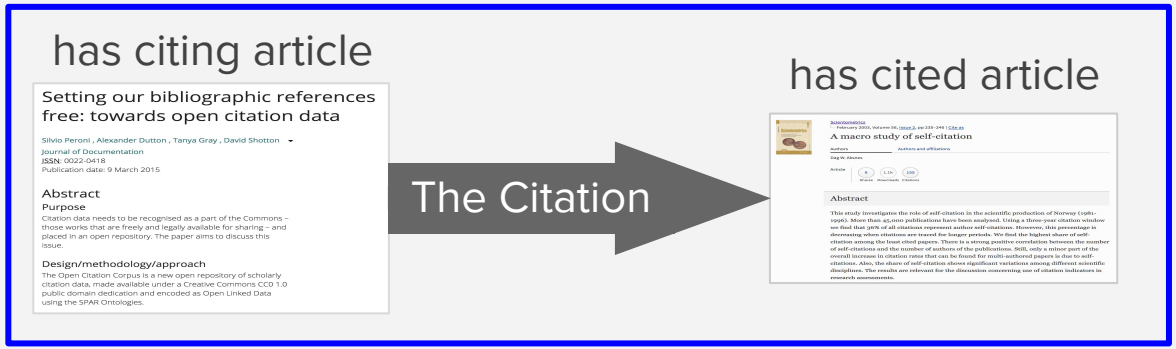


Citations as first-class data entities

Citations are normally treated simply as the links between published entities



Alternative richer view is to regard a **citation as a data entity** in its own right



Open Citation Identifier (OCI)

We defined the **Open Citation Identifier (OCI)**, a persistent identifier scheme for citations contained in bibliographic databases

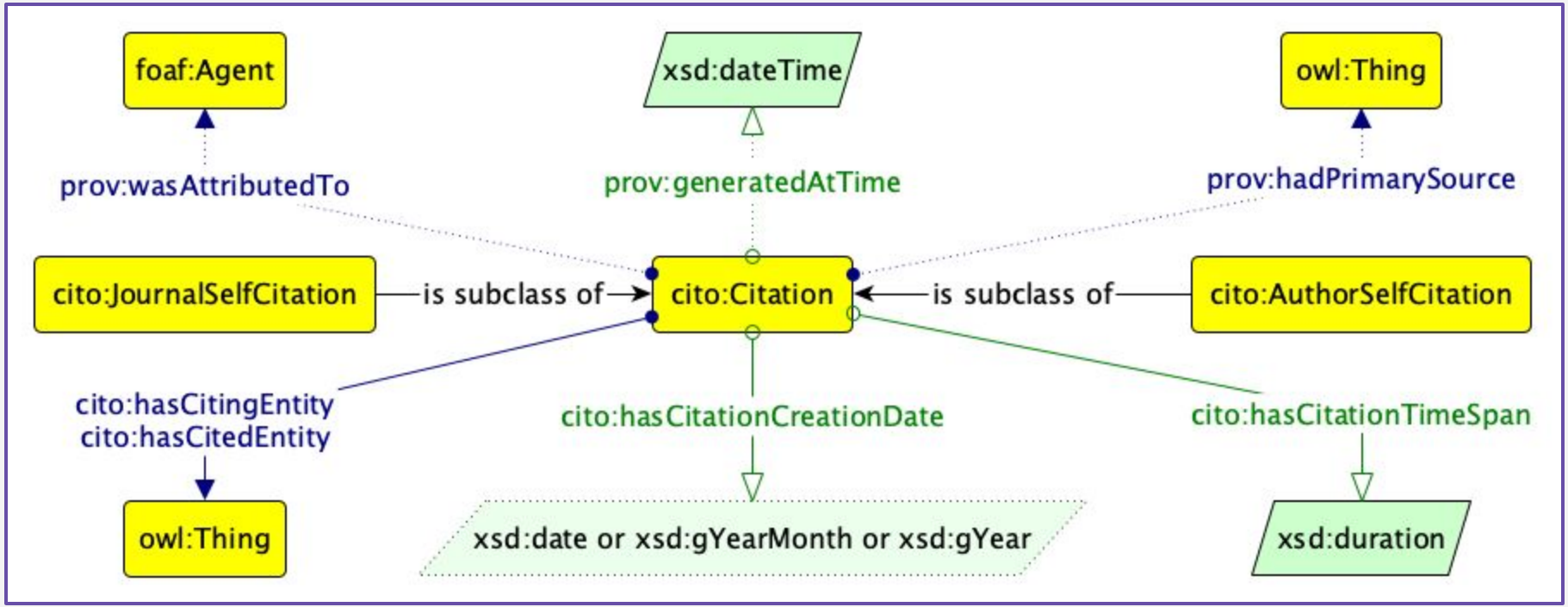
Structure: **oci:number-number,** where “oci:” is the identifier prefix

Some examples:

- *oci:01027931310-01022252312* (citation in Wikidata, identified by “010”)
- *oci:020010108063601070506630807020263066305 09-020010108063601070506630807020263056303 01* (citation in Crossref, identified by “020”)
- *oci:0302544384-0307295288* (citation in the OCC, identified by the “030”)

Peroni S, Shotton D (2019). Open Citation Identifier: Definition. Figshare. DOI: https://doi.org/10.6084/m9.figshare.7127816

OpenCitations Indexes sub-module



- Based mainly on the Citation Typing Ontology (CiTO) to define citations as **first-class data entities**.

Datasets

- **The OpenCitations Corpus** (OCC, http://opencitations.net/corpus): new instance was set up at the University of Bologna in early July 2016, and currently contains almost **14M citation links to over 7.5M cited resources**. OCC includes **information about different kinds of bibliographic entities** such as: bibliographic resources (e.g. journals, and journal articles), identifiers (local), responsible agents.
- **The OpenCitations Indexes** (http://opencitations.net/index): bibliographic indexes recording citations between publications using the data available in particular bibliographic databases.
    - **COCI** (launch: July 2018): ~445M citations between ~46M bibliographic entities
    - **CROCI** (launch: March 2019): 76 citations between 81 bibliographic entities

COCI

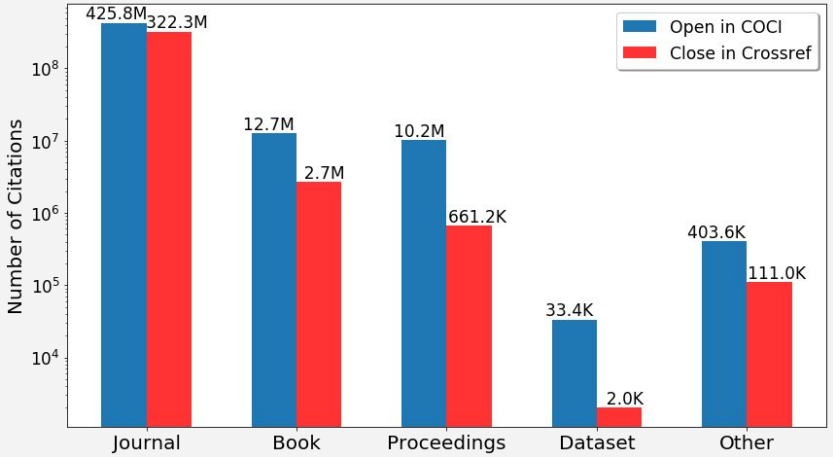
- **COCI,** the OpenCitations Index of Crossref open DOI-to-DOI citations (w3id.org/oc/index/coci).
- The first of the OpenCitations indexes. (w3id.org/oc/index), in which **citations are exposed as first-class data entities** with accompanying properties

Publisher	Outgoing citations	Incoming citations
Springer Nature	79,860,827	52,257,862
Wiley	76,819,685	48,174,542
Elsevier	2,853,739	96,310,027
Informa UK Limited	41,433,917	14,975,989
Institute of Electrical and Electronics Engineers (IEEE)	30,114,985	20,940,703
American Physical Society (APS)	15,729,297	16,065,862
SAGE Publications	15,933,805	7,915,082
Ovid Technologies (Wolters Kluwer Health)	9,971,274	12,840,293
Oxford University Press (OUP)	9,891,000	11,466,659
AIP Publishing	10,130,022	8,455,097

Heibi I, Peroni S, Shotton D (2019). COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations. arXiv:1904.06052

CROCI

- **CROCI,** the Crowdsourced Open Citations Index.
- **Why CROCI?**
  Analysis of COCI + Crossref shows that number of closed citations is of great significance



Publisher of entities receiving citations	Open citations in COCI	Close citations in Crossref
Elsevier BV	97,079,715 (47.92%)	105,486,201 (52.08%)
Springer Nature	52,655,655 (61.05%)	33,596,285 (38.95%)
Wiley	48,228,581 (56.61%)	36,970,208 (43.39%)
Institute of Electrical and Electronics Engineers (IEEE)	21,084,872 (86.21%)	3,373,087 (13.79%)
American Physical Society (APS)	16,211,918 (72.52%)	6,142,167 (27.48%)
American Chemical Society (ACS)	15,706,062 (42.28%)	21,438,766 (57.72%)
Informa UK Limited	15,066,947 (68.39%)	6,965,166 (31.61%)
Ovid Technologies (Wolters Kluwer Health)	12,903,492 (56.38%)	9,981,473 (43.62%)
Oxford University Press (OUP)	11,530,527 (62.95%)	6,785,205 (37.05%)
AIP Publishing	8,736,352 (64.05%)	4,904,576 (35.95%)
SAGE Publications	7,978,522 (73.31%)	2,905,035 (26.69%)
JSTOR	6,426,926 (61.50%)	4,023,765 (38.50%)
University of Chicago Press	5,600,609 (74.20%)	1,947,231 (25.80%)
IOP Publishing	5,412,557 (68.82%)	2,452,803 (31.18%)
American Association for the Advancement of Science (AAAS)	5,204,267 (54.34%)	4,372,753 (45.66%)
Proceedings of the National Academy of Sciences	5,046,601 (56.57%)	3,874,616 (43.43%)
Royal Society of Chemistry (RSC)	4,960,056 (49.32%)	5,096,291 (50.68%)
Cambridge University Press (CUP)	4,883,890 (68.85%)	2,209,466 (31.15%)
American Psychological Association (APA)	4,530,463 (66.67%)	2,264,755 (33.33%)

Publisher submitting references to Crossref	Closed
Elsevier BV	11,020,314
Institute of Electrical and Electronics Engineers (IEEE)	3,331,913
American Chemical Society (ACS)	496,855
University of Chicago Press	41,566

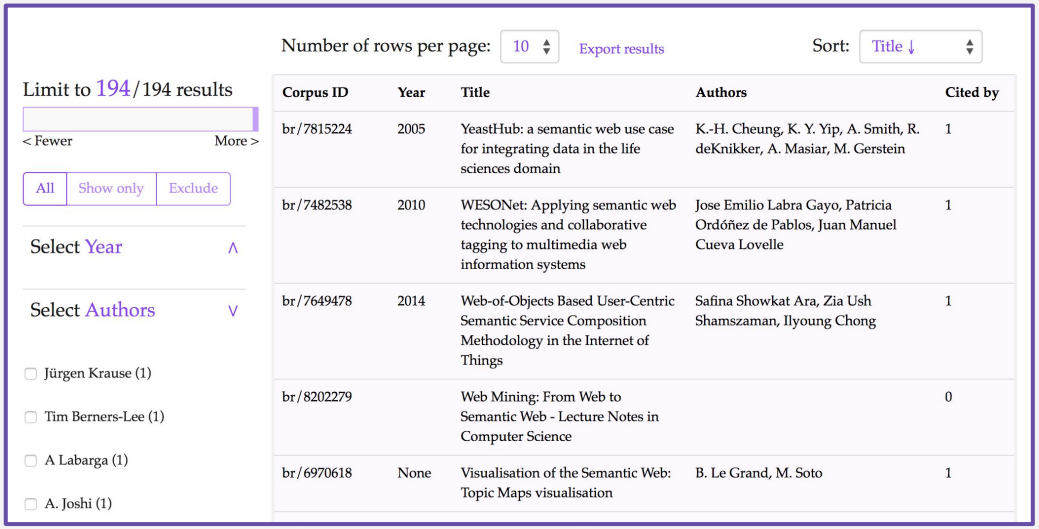
Heibi I, Peroni S, Shotton D (2019). Crowdsourcing open citations with CROCI - An analysis of the current status of open citations, and a proposal. arXiv:1902.02534

Online services

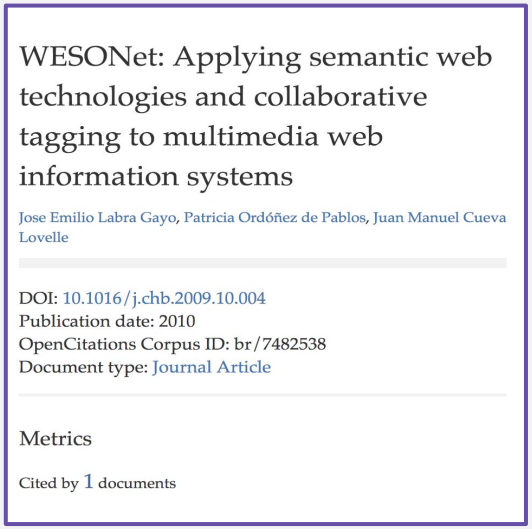
- Dumps:
  - **Corpus and Indexes data and provenance**
  - Available in **CSV, N-Triples** formats
- SPARQL endpoints:
  - **SPARQL: is the RDF Query Language**
  - Used to **Query** both the Corpus and Indexes
  - **Suitable for Semantic Web experts**
- REST APIs:
  - Implemented by means of **RAMOSE**, the Restful API Manager Over SPARQL Endpoints (https://github.com/opencitations/ramose).
  - Accessing/Retrieving the data of the Corpus and Indexes
    - Corpus: opencitations.net/api/v1
    - Indexes: opencitations.net/index/api/v1

Online services (Interface)

- Searching and Browsing interface:
  - developed through the two generic tools OSCAR and LUCINDA
  - Suitable for the common **non Semantic Web experts users**

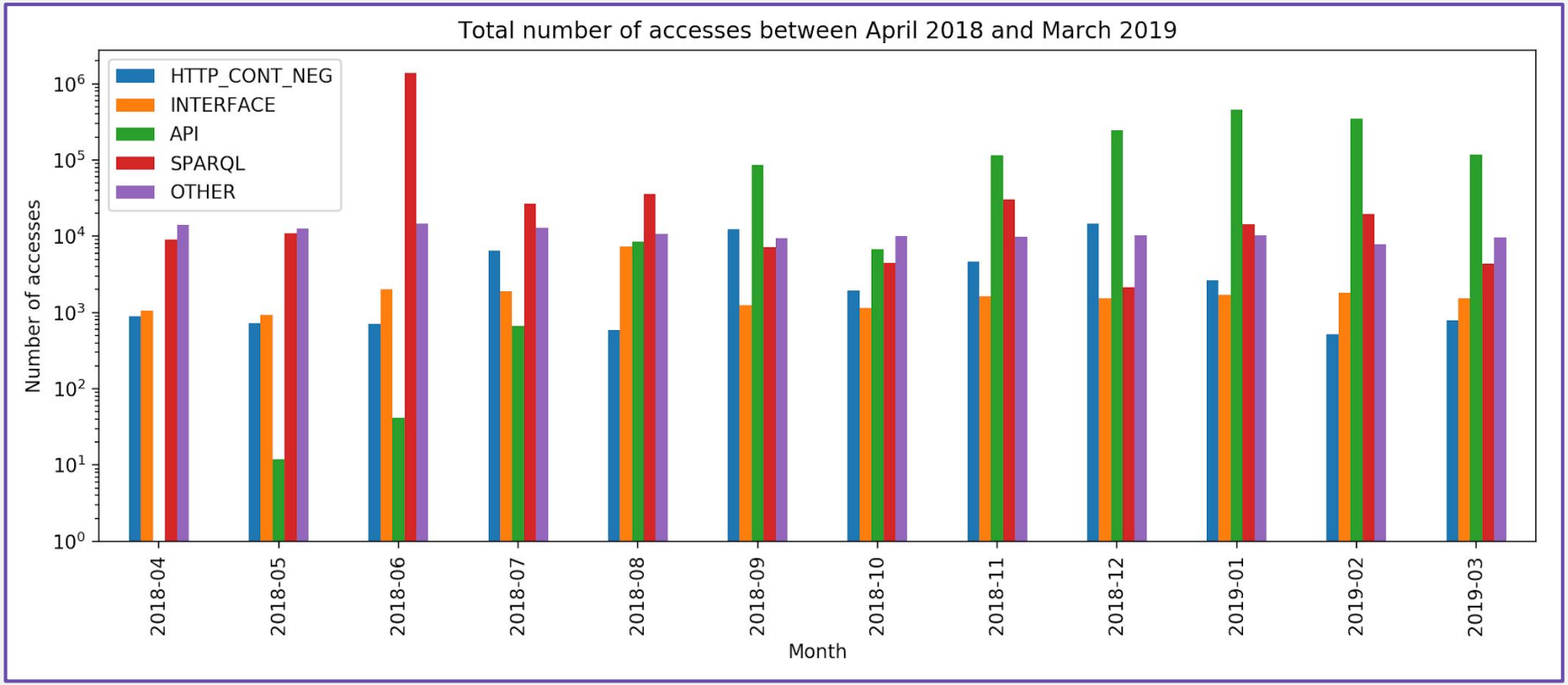


http://opencitations.net/search



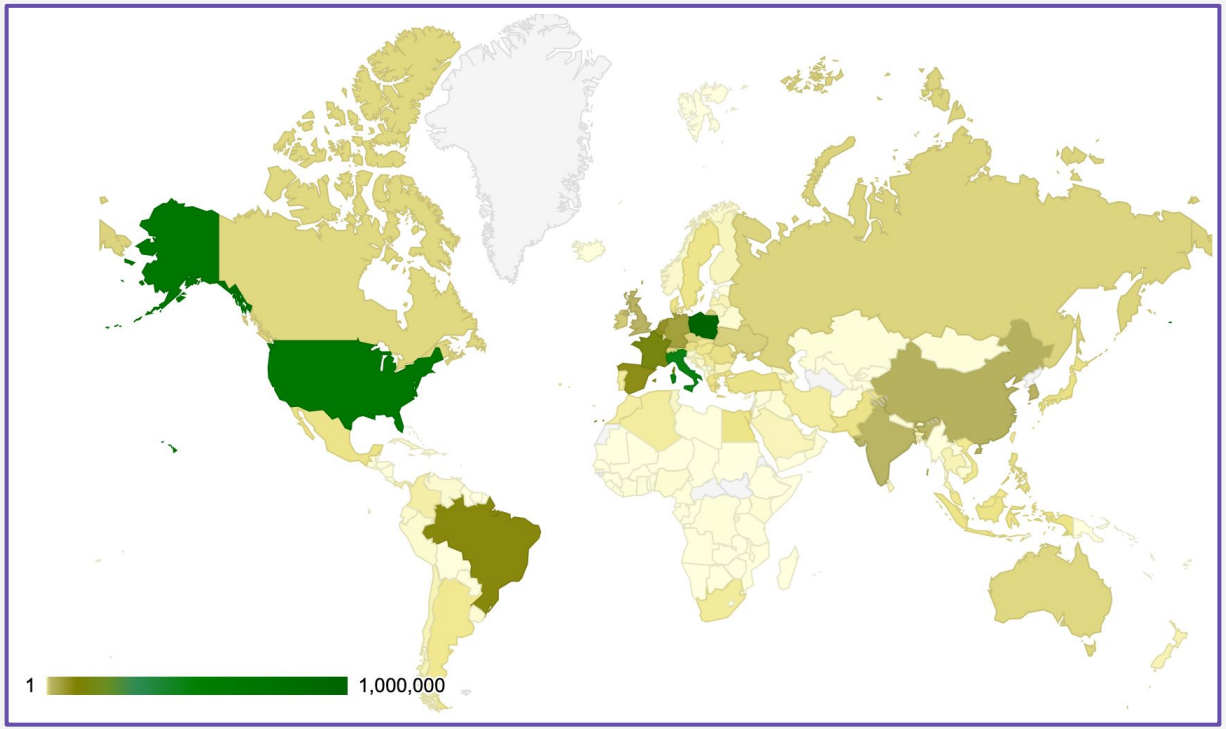
https://opencitations.net/browser/br/7482538

Services usage



Services usage: a map overview

Number of accesses to the OpenCitations services and website per country



Thank you for your attention

OpenCitations

an infrastructure for open citation data

@opencitations – http://opencitations.net

Ivan Heibi

Digital Humanities Advanced Research Centre (DHARC),
Department of Classical Philology and Italian Studies,
University of Bologna, Bologna (Italy)
ivan.heibi2@unibo.it – @ivanheib – https://ivanhb.github.io

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA