One year of the OpenCitations Corpus

Releasing RDF-based scholarly citation data into the Public Domain

Paper (HTML): https://w3id.org/people/essepuntato/papers/oc-iswc2017.html
Paper (DOI): https://doi.org/10.1007/978-3-319-68204-4_19
Slides (HTML): https://w3id.org/people/essepuntato/presentations/oc-iswc2017.html
Slides (DOI): https://doi.org/10.6084/m9.figshare.5526967

Silvio Peroni ✉ iD • David Shotton • Fabio Vitali

16th International Semantic Web Conference (ISWC 2017), 21-25 October 2017, Vienna, Austria

OpenCitations

Citations are the links that knit together our scientific and cultural knowledge

However often they are enclosed in closed citation indexes thought only for human consumption and/or accessible only by paying significant fees

The OpenCitations Corpus (OCC, http://opencitations.net) is a LOD repository of CC0 citation data described with SPAR Ontologies (http://www.sparontologies.net)

It provides >11M citation links from ~260,000 citing articles to ~6M cited resources + provenance information

Ingestion workflow

We developed several scripts for implementing the ingestion workflow that populates the OpenCitations Corpus

All the software is available on the OpenCitations GitHub repository and released as open source code with the ISC License

These scripts implement a *live* and *iterative* process

- Live: it's working while I'm speaking. It doesn't sleep, never. It's like a sentient, relentless, fast zombie – watch out!
- Iterative: the ingestion workflow continuously calls several external APIs to obtain new reference lists and clean metadata of the citing and cited papers

External APIs

At present, all the reference lists are taken by processing the XML sources of the papers in the PubMed Central Open Access subset

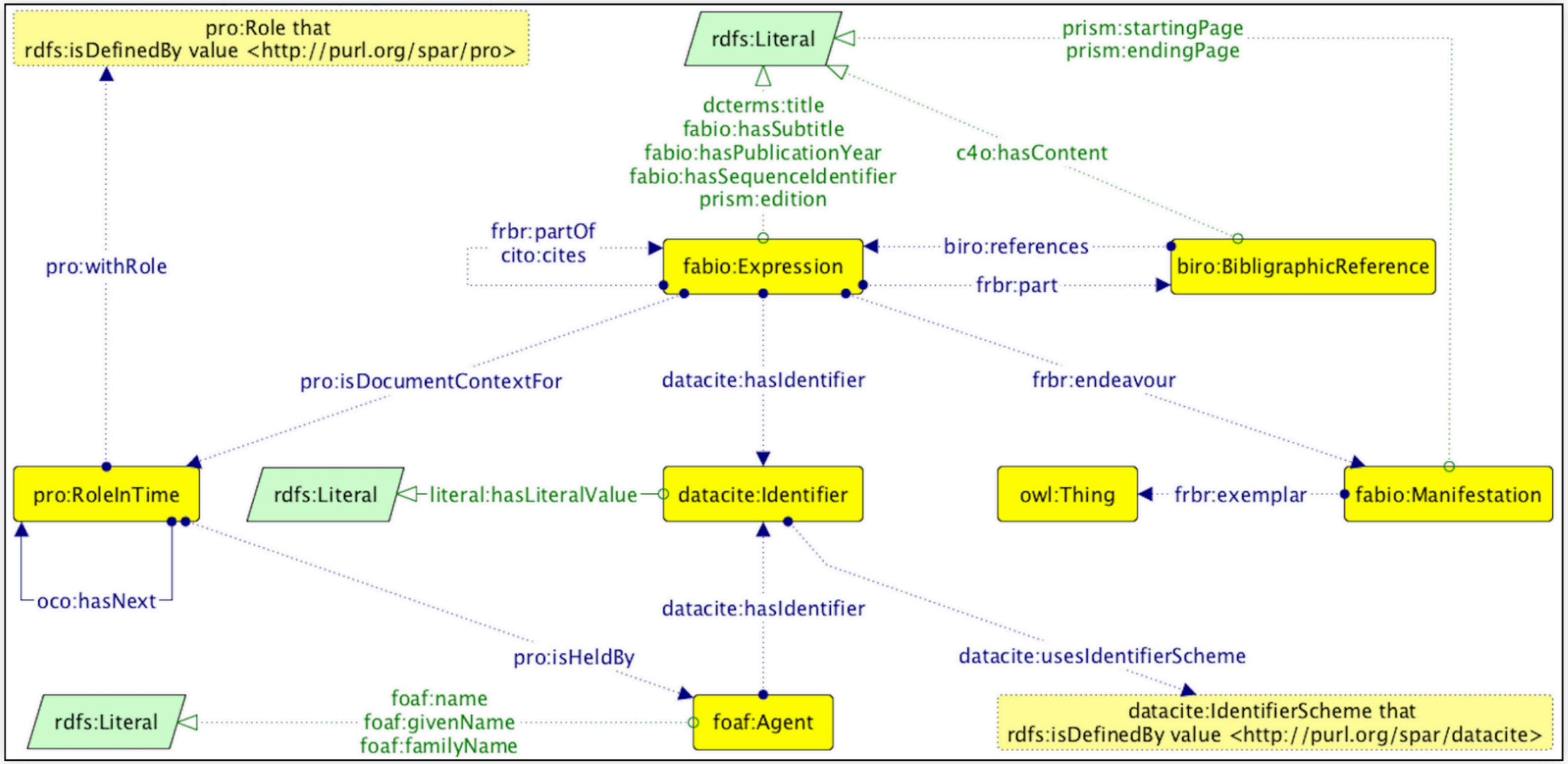
Europe PubMed Central API for retrieving the XML sources

- We ask for the most recent papers first
- Citing papers includes articles published in 2016 and 2017
- There are 1.75M OA articles available in PubMed, according to their API. We have harvested 15% so far...

Crossref APIs to obtain additional information (title, authors, venues, etc.) about citing/cited papers, and then call the ORCID APIs to obtain ORCIDs of the authors

Data model

Available at https://doi.org/10.6084/m9.figshare.3443876 and implemented in the OpenCitations Ontology



pro:Role that rdfs:isDefinedBy value <http://purl.org/spar/pro>

rdfs:Literal

prism:startingPage
prism:endingPage

dcterms:title
fabio:hasSubtitle
fabio:hasPublicationYear
fabio:hasSequenceIdentifier
prism:edition

c4o:hasContent

frbr:partOf
cito:cites

fabio:Expression

biro:references

frbr:part

biro:BibligraphicReference

pro:withRole

pro:isDocumentContextFor

datacite:hasIdentifier

frbr:endeavour

pro:RoleInTime

rdfs:Literal

literal:hasLiteralValue

datacite:Identifier

owl:Thing

frbr:exemplar

fabio:Manifestation

oco:hasNext

datacite:hasIdentifier

pro:isHeldBy

datacite:usesIdentifierScheme

rdfs:Literal

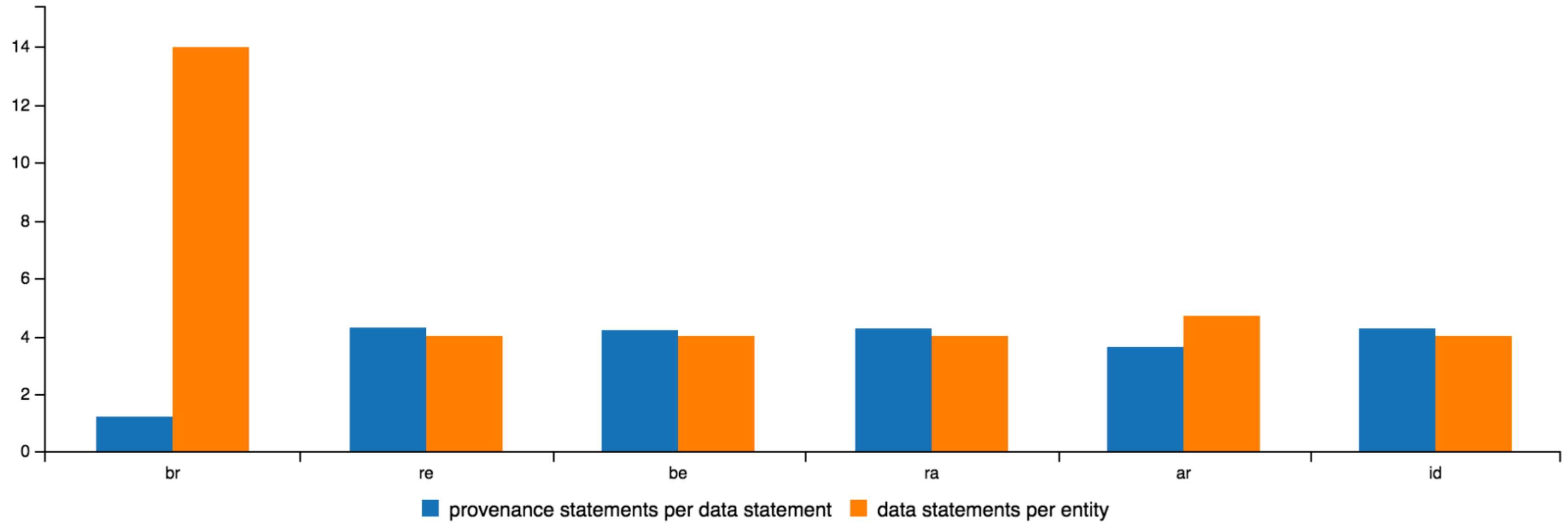
foaf:name
foaf:givenName
foaf:familyName

foaf:Agent

datacite:IdentifierScheme that rdfs:isDefinedBy value <http://purl.org/spar/datacite>

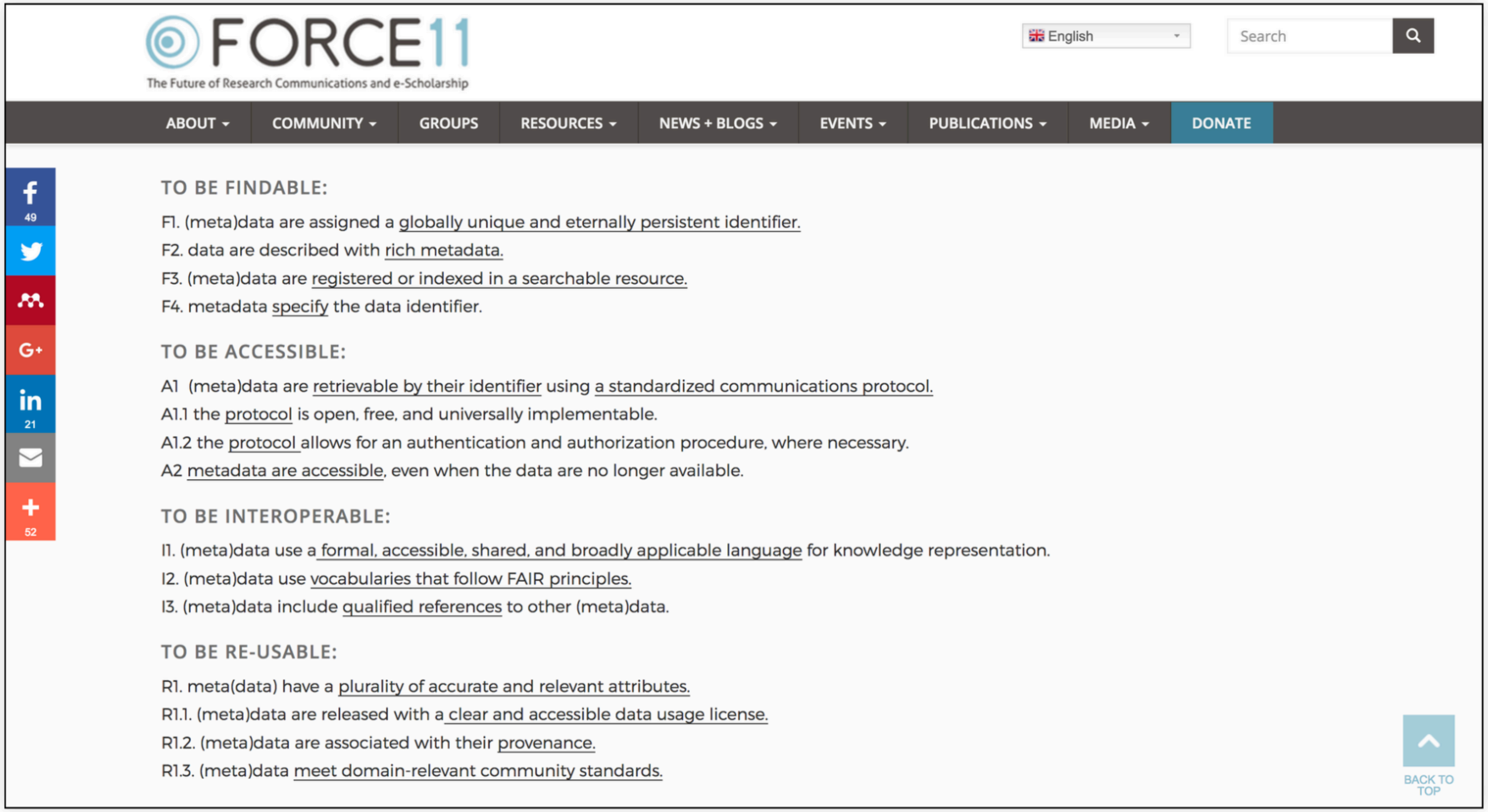
Resources in the OCC

Entity type	What it describes	Count in the OCC
Bibliographic resource (br)	Conference papers, book chapters, journal articles, academic proceedings, books, journals, etc.	~7.3M
Resource embodiment (re)	Digital vs. print, first and ending pages, etc.	~4.3M
Bibliographic entry (be)	Textual content of a reference in a reference list	~10.5M
Resposible agent (ra)	Given name, family name and ORCID of the agent involved	~22.5M
Agent role (ar)	Author, publisher, etc.	~28.6M
Identifier (id)	DOI, PubMed ID, PubMed Central ID, ORCID, ISSN, etc.	~15.1M



Data statements	Provenance statements
~0.45B	~1.5B

FAIR citation data



TO BE FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:

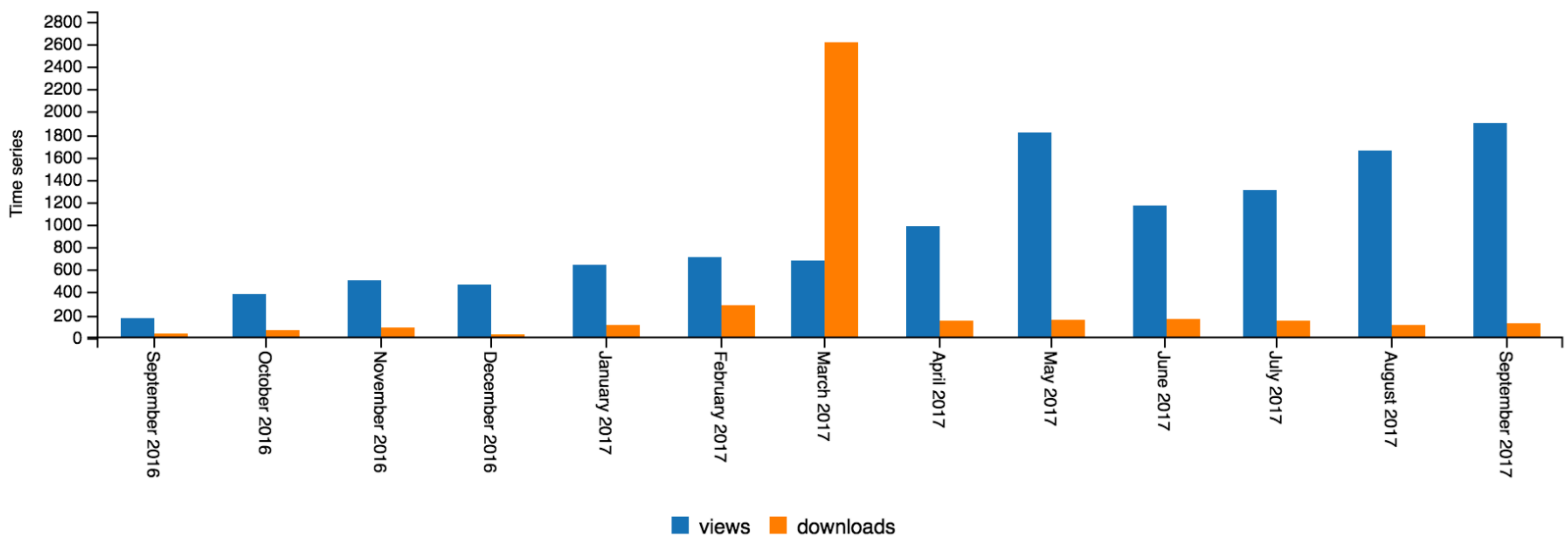
R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

More details: http://w3id.org/people/essepuntato/papers/oc-garr2017.html

Accessing OCC data

- Direct access to bibliographic resources by means of their HTTP URIs (via content negotiation, e.g. https://w3id.org/oc/corpus/br/1)
- SPARQL endpoint (https://w3id.org/oc/sparql)
- Monthly dumps (http://opencitations.net/download, stored in Figshare)

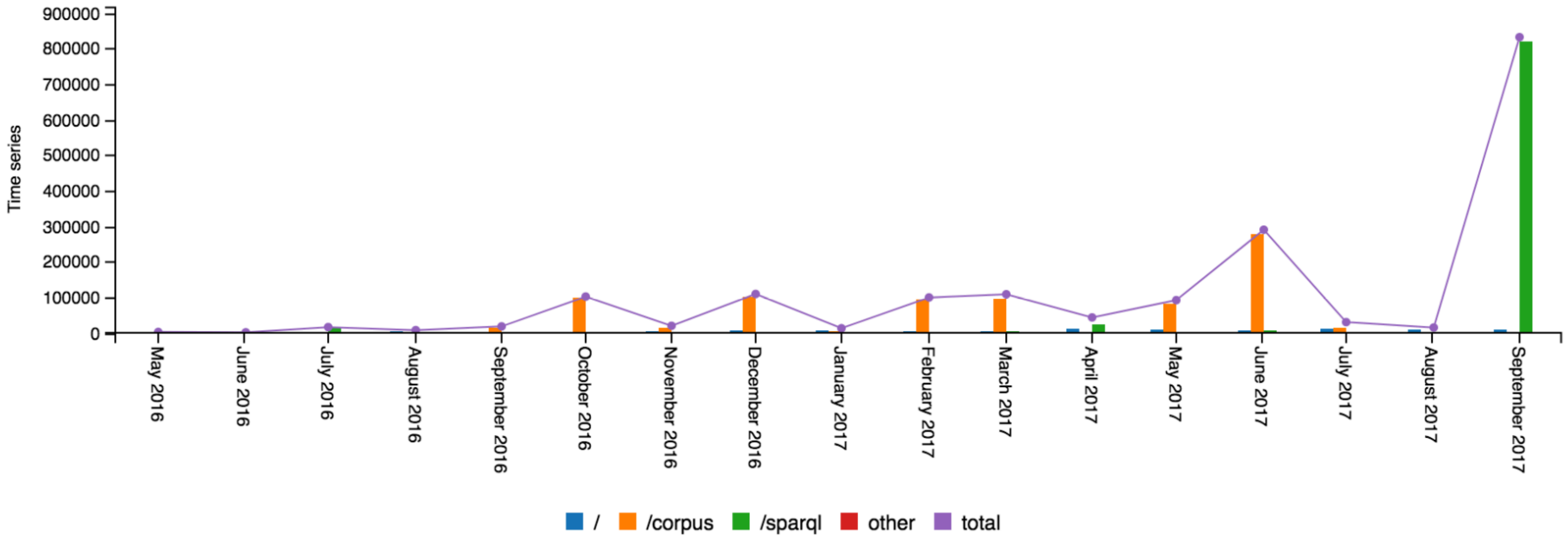
Dump views and downloads



Usage of the OCC

Some known adopters: Wikidata, OpenAIRE, LOC-DB, eLife, Ontotext, independent researchers (Anna Kamińska, Daniel Himmelstein, Thiago Nunes and Daniel Schwabe)

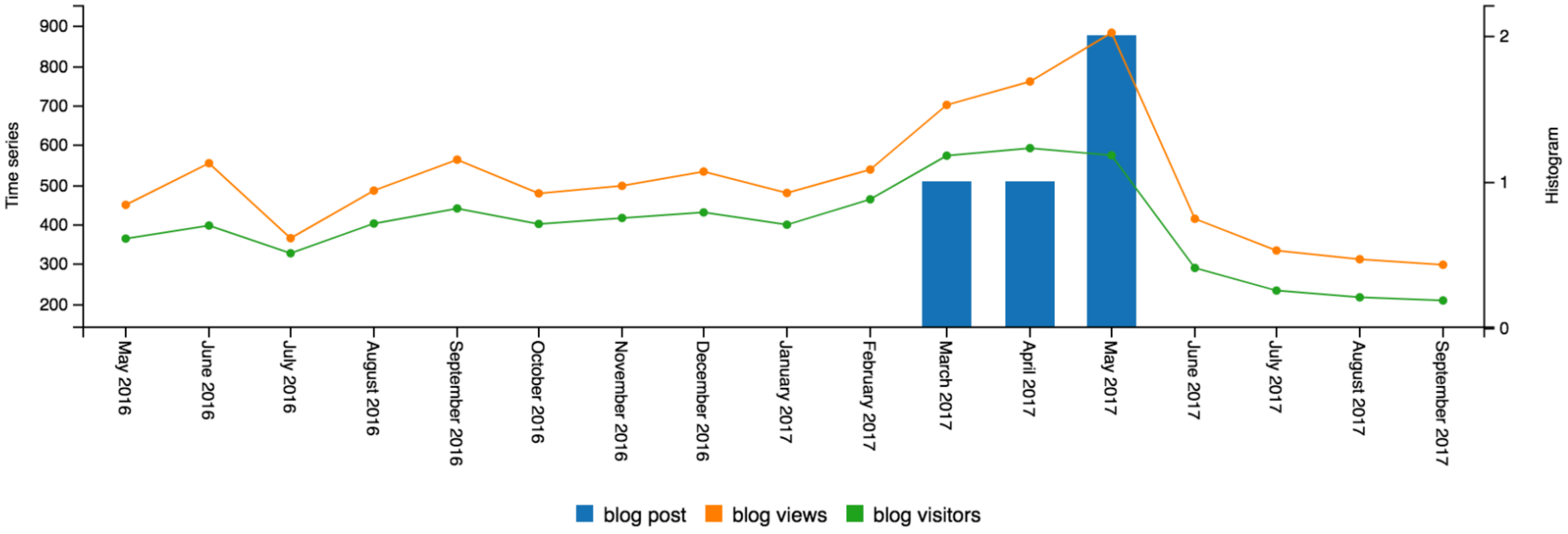
Website accesses



Use by country: United States (33.7%), France (31.3%), Bulgaria (14.3%)

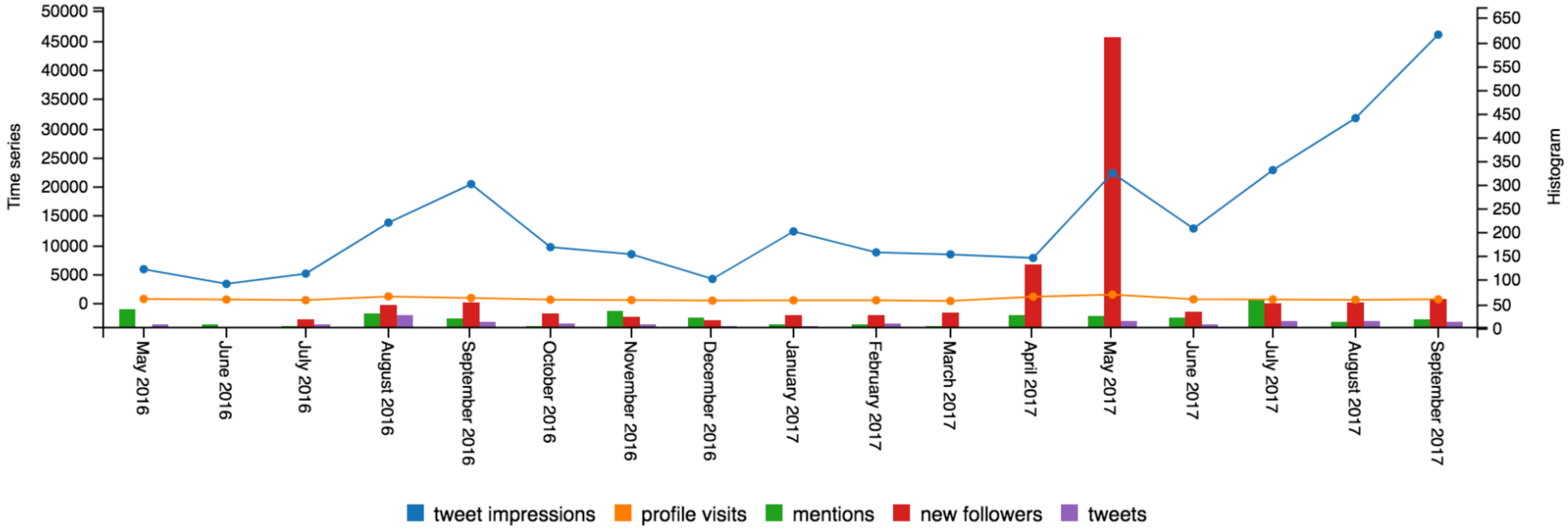
Blog

OpenCitations Blog on Wordpress at https://opencitations.wordpress.com

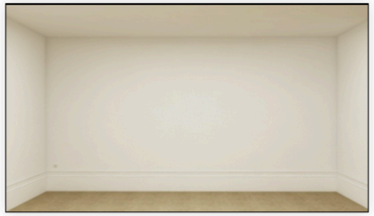


Twitter

OpenCitations Twitter account at https://twitter.com/opencitations



Current and future infrastructure



It's a virtual machine! It doesn't exist

DELL PowerEdge R730xd V4, 512GB RAM, 22TB HD + 30 Raspberry Pi 3

We have recently received a small grant from the Sloan Foundation for one year's salary for a postdoc to develop new user interfaces, and new hardware to enhance the OCC performance - from 8M citations per year to **240M**

Data are coming

END

One year of the OpenCitations Corpus

Silvio Peroni ✉ iD 🐦 • David Shotton • Fabio Vitali

16th International Semantic Web Conference (ISWC 2017), 21-25 October 2017, Vienna, Austria