Linked Open Citation
Database Workshop
Schloss Mannheim
7 November 2017

OpenCitations

OpenCitations as a Hub for Open Citation Data

David Shotton

david.shotton@opencitations.net

Oxford e-Research Centre
University of Oxford, UK

The importance of citations

people get upset. When our project was announced through a press release, comments appeared on the website of the British National Party, a far-right British political organization that opposes immigration and favours 'voluntary repatriation'. One contributor described our project as: 'Another government-funded (sic) justification for immigration which will say there is no such thing as indigenous British', and our research group as 'a bunch of Marxist parasites who have never done a proper day's work'. Well, proper or not, there is much work to do, and we Marxist parasites have high hopes that it will produce new insights.

My work is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (grant no. 087576).

REFERENCES

1 1000 Genomes Project Consortium, Durbin, R. M. *et al.* 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073. (doi:10.1038/nature09534)
2 Barbujani, G. & Colonna, V. 2010 Human genome diversity: frequently asked questions. *Trends Genet.* **26**, 285–295. (doi:10.1016/j.tig.2010.04.002)
3 Mardis, E. R. 2008 Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402. (doi:10.1146/annurev.genom.9.081307.164359)
era. *Annu. Rev. Genomics Hum. Genet.* **12**, 245–274. (doi:10.1146/annurev-genom-090810-183123)
14 Underhill, P. A. & Kivisild, T. 2007 Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* **41**, 539–564. (doi:10.1146/annurev.genet.41.110306.130407)
15 Ragoussis, J. 2009 Genotyping technologies for genetic research. *Annu. Rev. Genomics Hum. Genet.* **10**, 117–133. (doi:10.1146/annurev-genom-082908-150116)
16 Pritchard, J. K., Stephens, M. & Donnelly, P. 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.
17 Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. & Feldman, M. W. 2002 Genetic structure of human populations. *Science* **298**, 2381–2385. (doi:10.1126/science.1078311)
18 Li, J. Z. *et al.* 2008 Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104. (doi:10.1126/science.1153717)
19 Novembre, J. *et al.* 2008 Genes mirror geography within Europe. *Nature* **456**, 98–101. (doi:10.1038/nature07331)
20 Lao, O. *et al.* 2008 Correlation between genetic and geographic structure in Europe. *Curr. Biol.* **18**, 1241–1248. (doi:10.1016/j.cub.2008.07.049)
21 Pool, J. E. & Nielsen, R. 2009 Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711–719. (doi:10.1534/genetics.108.098095)
22 Moorjani, P. *et al.* 2011 The history of African gene flow into Southern Europeans, Levantines, and Jews.

Making citations is an essential scholarly activity

- A citation is created by an author's performative act of citing a published work that is relevant to the current work, typically made by including a bibliographic reference in the reference list of the current work
- A citation is a permanent directional conceptual link from the citing bibliographic work to a cited work

Making citations is an essential scholarly activity

- A citation is created by an author's performative act of citing a published work that is relevant to the current work, typically made by including a bibliographic reference in the reference list of the current work
- A citation is a permanent directional conceptual link from the citing bibliographic work to a cited work

- It permits an author to give credit to another person's endeavours that have played a part in the development of the author's own ideas or results
- Direct citation is a key indicator of a publication's significance

Making citations is an essential scholarly activity

- A citation is created by an author's performative act of citing a published work that is relevant to the current work, typically made by including a bibliographic reference in the reference list of the current work
- A citation is a permanent directional conceptual link from the citing bibliographic work to a cited work

- It permits an author to give credit to another person's endeavours that have played a part in the development of the author's own ideas or results
- Direct citation is a key indicator of a publication's significance

- Citations also integrate our independent acts of scholarship into a global knowledge network
- Bibliometric analysis of the flow of information and ideas through the citation network, and its changes over time, can reveal patterns of communication between scholars and the development and demise of academic disciplines

How is the present situation imperfect?

- The present scholarly citation system inadequately exposes the knowledge networks that exist within the scholarly literature, linking papers, authors, funders, research projects and datasets
- Citation data are hidden behind subscription firewalls of commercial companies
- Academics are not free to use their own citation data as they please

How is the present situation imperfect?

- The present scholarly citation system inadequately exposes the knowledge networks that exist within the scholarly literature, linking papers, authors, funders, research projects and datasets
- Citation data are hidden behind subscription firewalls of commercial companies
- Academics are not free to use their own citation data as they please

- In this Open Access age, it is a scandal that reference lists from journal articles, the core elements of the academic data cycle, are not freely available for use by the scholars who created them

- Citation data now need to be recognized as a part of the Commons – those works that are freely and legally available for sharing
- The Initiative for Open Citations (I4OC) is working to achieve this

The OpenCitations Corpus

- OpenCitations (http://opencitations.net) is an infrastructure organization directed by myself and by Silvio Peroni of the University of Bologna
- Its primary purpose is to host and develop the OpenCitations Corpus (OCC), a Linked Open Data repository of bibliographic citation data covering all disciplines

The OpenCitations Corpus

- OpenCitations (http://opencitations.net) is an infrastructure organization directed by myself and by Silvio Peroni of the University of Bologna
- Its primary purpose is to host and develop the OpenCitations Corpus (OCC), a Linked Open Data repository of bibliographic citation data covering all disciplines

- The first OCC prototype was created at Oxford in 2011 with Jisc funding
- A new instance of the OCC, based on our revised OpenCitations Metadata Model, was then established by Silvio Peroni at the University of Bologna
- It has been ingesting scholarly references continuously since early July 2016

The OpenCitations Corpus

- OpenCitations (http://opencitations.net) is an infrastructure organization directed by by myself and Silvio Peroni of the University of Bologna
- Its primary purpose is to host and develop the OpenCitations Corpus (OCC), a Linked Open Data repository of bibliographic citation data covering all disciplines

- The first OCC prototype was created at Oxford in 2011 with Jisc funding
- A new instance of the OCC, based on our revised OpenCitations Metadata Model, was then established by Silvio Peroni at the University of Bologna
- It has been ingesting scholarly references continuously since early July 2016

- OCC now provides the largest RDF collection of open citation data on the Web
  - Currently holds references from ~270,000 citing bibliographic resources
  - **Provides >11 million citation links to over 6 million cited resources**
  - These data are freely available under a CC0 public domain waiver

Source data - reference lists from PubMed Central

- At present, the ingested reference lists are obtained by processing the XML sources of papers in the Open Access subset of PubMed Central
- These are parsed to yield authors, titles, journal names, etc.
  - We ask for the most recent papers first
  - Thus, as citing papers, the OCC mainly includes articles published in 2016 and 2017
- The identifiers of all the citing papers that have already processed are stored locally, so as not to request the same XML source twice
- We then call several external APIs, including Crossref and ORCID, to obtain additional metadata describing the citing and cited papers and their authors

- There are almost 1.7 million OA articles available in PubMed Central
  - So far we have harvested 16% . . .

The raw reference list data

- The reference lists extracted from citing papers are made available in JSON:


```
{
  "doi": "10.1007/s11892-016-0752-4",
  "pmid": "27168063",
  "pmcid": "PMC4863913",
  "localid": "MED-27168063",
  "curator": "BEE EuropeanPubMedCentralProcessor",
  "source": "http://www.ebi.ac.uk/europepmc/webservices/rest/PMC4863913/fullTextXML",
  "source_provider": "Europe PubMed Central”
  "references": [
    ...
```


The citing paper's metadata and identifiers


```
  {
    "bibentry": "Chang, KY, Unanue, ER. Prediction of HLA-DQ8beta cell peptidome using
                 a computational program and its relationship to autoreactive T cells,
                 Int Immunol, 2009, 21, 6, 705, 13, DOI: 10.1093/intimm/dxp039,
                 PMID: 19461125",
    "pmid": "19461125",
    "doi": "10.1093/intimm/dxp039",
    "pmcid": "PMC2686615",
    "process_entry": "True”
  },
  ...
]
}
```


A reference in the citing paper's reference list, with its own ids

The SPAR (Semantic Publishing and Referencing) Ontologies

- OCC data are then stored in RDF (JSON-LD) using the SPAR (Semantic Publishing and Referencing) ontologies and other standard vocabularies

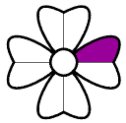
SPAR 

http://www.sparontologies.net/

- These SPAR ontologies include

FaBiO, the FRBR-aligned Bibliographic Ontology - an ontology for describing bibliographic entities (books, articles, etc.)

CiTO, the Citation Typing Ontology - an ontology that enables the characterization of citations, both factually and rhetorically



BiRO, the Bibliographic Reference Ontology - an ontology to define bibliographic records and references, and their compilation into bibliographic collections and reference lists, respectively

Availability of the OpenCitations Corpus data

- The data in the OpenCitations Corpus are available in three different ways:
  - Direct access to bibliographic resources by means of their HTTP URIs (via content negotiation), e.g. https://w3id.org/oc/corpus/br/1
  - Queries to our SPARQL endpoint: https://w3id.org/oc/sparql
  - Monthly dumps stored in Figshare: http://opencitations.net/download
- All the OpenCitations software is available on GitHub under an open license

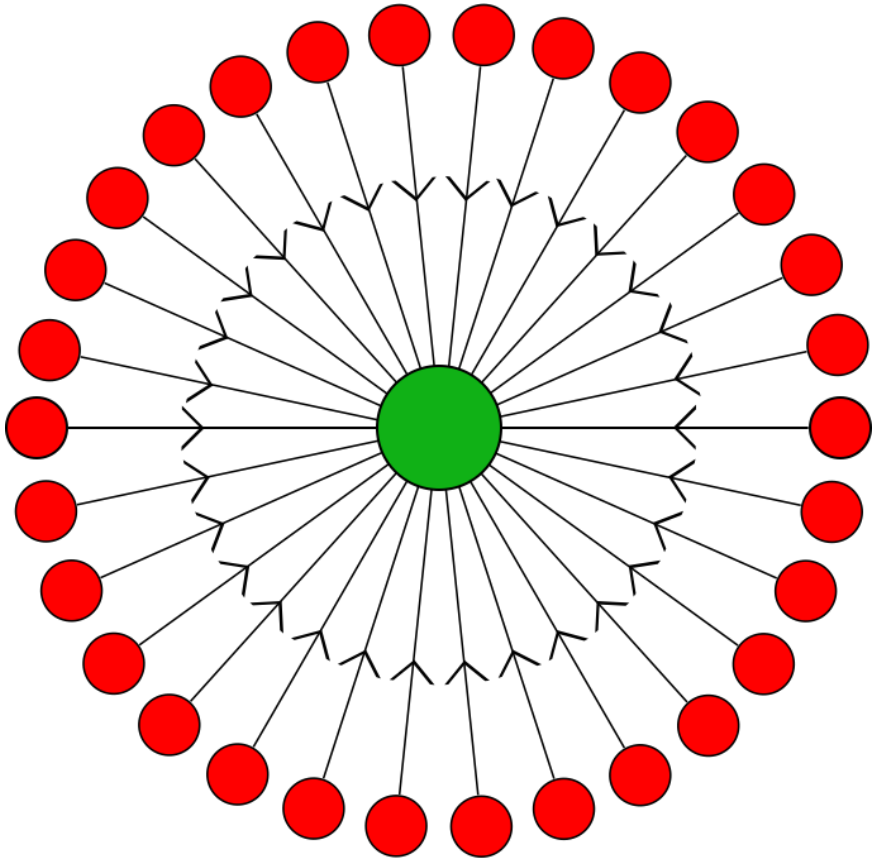
Availability of the OpenCitations Corpus data

- The data in the OpenCitations Corpus are available in three different ways:
  - Direct access to bibliographic resources by means of their HTTP URIs (via content negotiation), e.g. https://w3id.org/oc/corpus/br/1
  - Queries to our SPARQL endpoint: https://w3id.org/oc/sparql
  - Monthly dumps stored in Figshare: http://opencitations.net/download
- All the OpenCitations software is available on GitHub under an open license

- Currently the OCC uses a good graph-based triplestore – Blazegraph
- However, the virtual machine that hosts it is very limited in resources, causing performance problems for demanding SPARQL queries

The OpenCitations Enhancement Project

- We have recently received a grant from the Sloan Foundation for the OpenCitations Enhancement Project
  - This provides salary for a postdoc, just appointed, to develop new user interfaces for exploring the citations
  - And new hardware, presently being commissioned, to enhance the OCC performance
- This will use 30 Raspberry Pis harvesting reference lists in parallel
- These will feed JSON files to a central physical server powerful enough to manage complex SPARQL queries
- The ingestion rate will increase ~30-fold



The OpenCitations ingestion rate

- OpenCitations current ingests ~8 million new citations per year
- With our new hardware, this rate will increase to ~240 million new citations per year
- By the end of 2018, OpenCitations should hold ~ 250 million citations, compared to Web of Knowledge's ~1.25 billion

- **Even this partial coverage will include citations of all important papers**, these critical papers being easily recognized because they are highly cited, forming nodes in the citation graph with a large number of inward citation links

- A further five-fold increase in ingest rate - significant but achievable with additional hardware (and funding!) - would enable us to reach parity by 2020

Where will the references come from?

- We will quickly consume all 1.7 million articles in the Open Access Subset of PubMed Central
- We will then start harvesting the references from the ~16 million articles already made open at Crossref in response to The Initiative for Open Citations (I4OC, https://i4oc.org/), of which OpenCitations is a founding member
- Possible additional significant sources of open citation data include
  - ArXiv (1.3 million preprints, mainly in physics and the hard sciences)
  - CiteSeerX (>120 million references from >6 million documents)
  - CitEc (11 million references from a million Economics papers)
- References from pre-digital publications extracted by text mining, e.g.
  - In the Social Sciences, from the LOC-DB at the University of Mannheim
  - In Biological Taxonomy, mined into BioStor by Rod Page from the Biodiversity Heritage Library, e.g. http://biostor.org/reference/105357

Other users of OpenCitations

- Wikidata is pulling OpenCitations citation data to enrich Wikidata pages
- OpenAIRE is using OpenCitations bibliographic resource info in OpenAIRE
- Ludo Waltmann (Leiden Univ) is working with us to extend his VOSviewer software for exploring citation graphs to work with OpenCitations RDF data
- Tomas Petricek (Turing Institute) is extending his Gamma Project visualization software to handle OpenCitations RDF data
- Ontotext.com is demonstrating SPARQL query federation by combining Springer's SciGraph data with OpenCitations data
- Anna Kamińska (Polish Librarians Association) is undertaking bibliometric citation network analysis of PLoS One research papers using data in the OCC
- Daniel Schwabe, (Pontifical Catholic University of Rio de Janeiro), who has developed a tool *Xplain* for navigating a huge dataset of RDF relations, is using the OpenCitations Corpus as exemplar dataset
- . . . there are almost certainly others that we don't know about!

Adopting the OpenCitations Data Model

- The OpenCitations data model provides the possibility of interoperability between independent citation collections
- Other organizations and projects have adopted, or are considering adoption of the OpenCitations data model
  - LOC-DB has adopted the OpenCitations Data Model
  - Erick Peirson, the lead software architect at arXiv.org, has recently written to me asking for advice about structuring and publishing all arXiv references, and I have recommended adoption of the OpenCitations Model
  - Matteo Romaniello (Lausanne), who is storing citations of classical texts in Venetian scholarly documents as RDF, is also considering this
- At OpenCitations we are exploring the possibility of OCC mirror sites around the world, all using the same data model
- In this way, we hope that OpenCitations can become a hub for open citation data

Citations as First Class Data Objects – extensions of CiTO

- Conventionally, we think of citations just as the links between bibliographic resources, with the emphasis placed on those citing and cited papers
- However, citations are increasingly important in their own right, for example in the calculation of various metrics
- To make citations easier to deal with programmatically, I want to promote citations to become First Class Data Objects, with their own identifiers and properties
- For this purpose, we have added the following classes and properties to CiTO, the Citation Typing Ontology:
  - Class — ***cito:Citation***
    - with sub-class — ***cito:SelfCitation***
  - Data property — ***cito:hasCitationTimeSpan***

Citations as First Class Data Objects – CitationIDs

- Within the OpenCitations Data Model, we have also defined a new citation identifier, the **CitationID**
- To avoid duplication of information within the OpenCitations Corpus triplestore, these CitationIDs are created on the fly as and when required
  - For that purpose, CitationIDs have ‘virtual’ URIs
- Consider a citation recorded within the OpenCitations Corpus from Bibliographic Resource br/171 to Bibliographic Resource br/1048
  - Each BR has a URI of the form https://w3id.org/oc/corpus/br/171
- The citation itself will have an OCC internal CitationID ‘171-1048’
  - and its URI will be https://w3id.org/oc/virtual/citation/171-1048
- We hope that these additions to CiTO and to the OpenCitations Data Model will make it easier to specify and describe individual citations in RDF, and to process them automatically

Thank you!

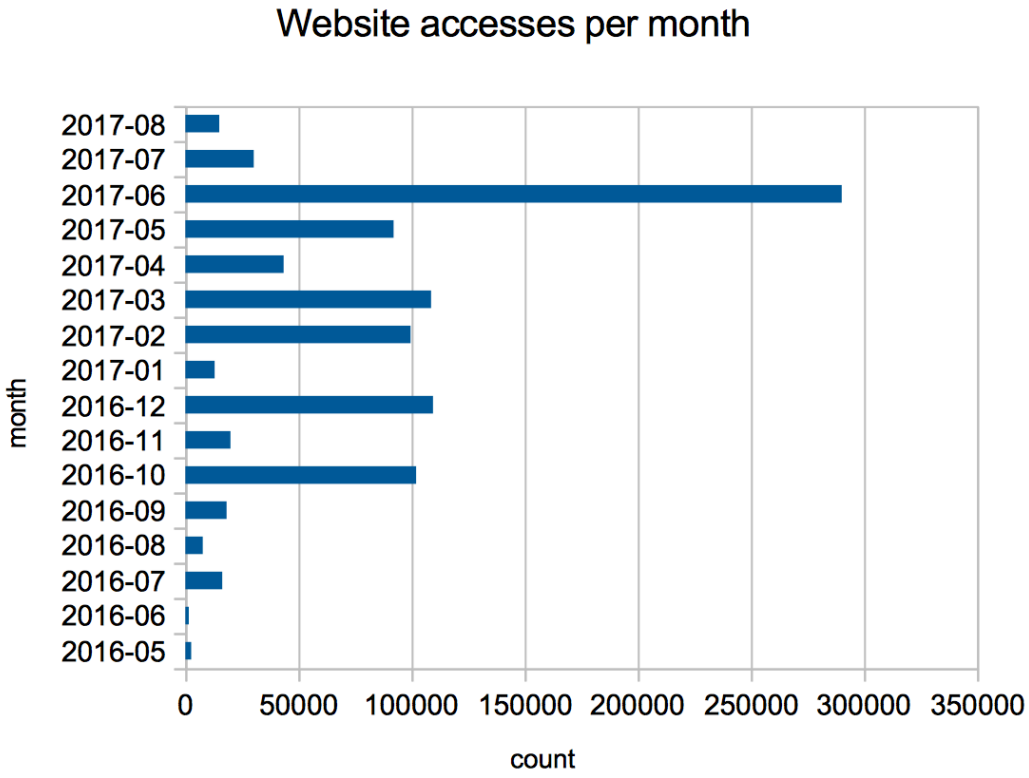
David Shotton
david.shotton@opencitations.net

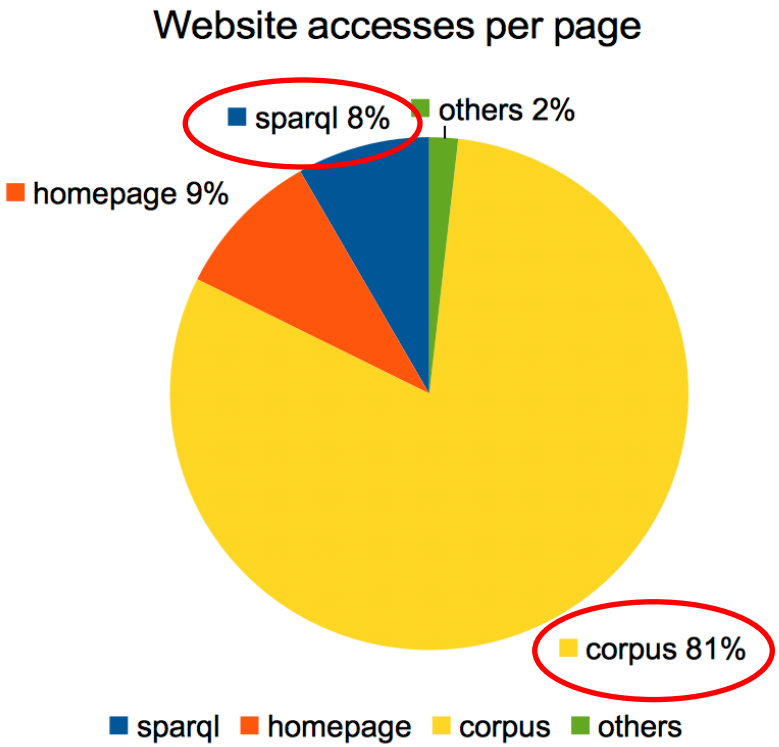
Silvio Peroni
silvio.peroni@opencitations.net

Website: http://opencitations.net
Email: contact@opencitations.net
Twitter: @opencitations
Blog: https://opencitations.wordpress.com

Use of the OpenCitations web site

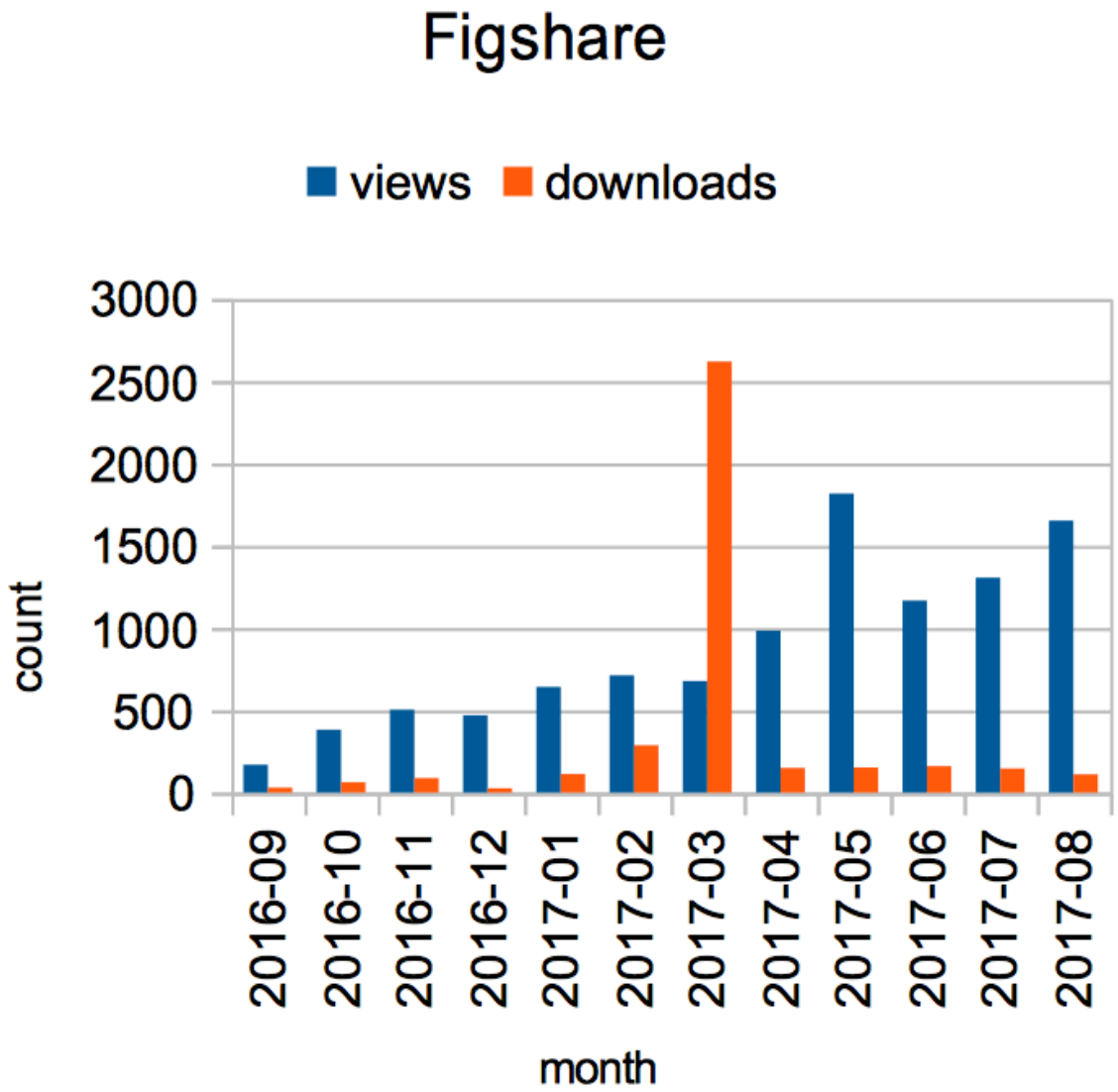
- Accesses to the OpenCitations web site and services:





The "corpus" and "sparql" pages have together gained 89% of the total accesses, showing that people mainly access the OpenCitations Corpus to explore and use the data within it

Use of OpenCitations data dumps on Figshare



What happened this summer?

- Use of the OpenCitations social accounts
  - Twitter - https://twitter.com/opencitations
  - Wordpress Blog – https://opencitations.wordpress.com/

increased markedly following the launch of the Initiative for Open Citations