**Workshop on Open Citations**
**2018**
University of Bologna
Bologna, Italy
3 September 2018

OpenCitations

# Citations in the wild, and how we are taming them

David Shotton

david.shotton@opencitations.net

Oxford e-Research Centre
University of Oxford, UK

# Cautionary tale #1: Shortcomings of current citation indexes

- Academics, university administrators and funders use Web of Science and the other commercial citation indexes not because they are perfect, but because they are the best we've got, and the data they supply are of real value
- One of the most important metrics they provide is the citation count, a proxy for the 'value' of the cited paper
- However, citation counts for individual papers vary wildly between indexes, as I will now demonstrate from my own publications

# Cautionary tale #1: Shortcomings of current citation indexes

- Academics, university administrators and funders use Web of Science and the other commercial citation indexes not because they are perfect, but because they are the best we've got, and the data they supply are of real value
- One of the most important metrics they provide is the citation count, a proxy for the 'value' of the cited paper
- However, citation counts for individual papers vary wildly between indexes, as I will now demonstrate from my own publications
- I took 25 of my most significant publications (not necessarily highly cited)
  - Publication dates evenly distributed across 46 years, 1970 to 2016
- For 20 of these, citation counts are given by all four indexes:
  - Web of Science, Scopus, Google Scholar and Microsoft Academic
  - If Dimensions and Semantic Scholar are included, that number drops to 16
- These fall into two topic categories
  - 13 pre-2000 biological papers and reviews: Molecular and cell biology, and microscopy (journal articles and book chapters)
  - 7 post-2000 computer science papers: Semantic web and open citations (journal articles and conference proceedings)

# Citation counts for the 13 pre-2000 biological papers

| WoS | Scopus | Google Scholar | Microsoft Academic | Mean | SD | CV |
|---|---|---|---|---|---|---|
| 49 | 40 | 46 | 58 | 48.3 | 7.5 | 16% |
| 250 | 189 | 269 | 275 | 246 | 39.3 | 16% |
| 20 | 51 | 26 | 54 | 37.8 | 17.3 | 46% |
| 50 | 35 | 51 | 57 | 48.3 | 9.4 | 19% |
| 533 | 346 | 515 | 524 | 480 | 89.3 | 19% |
| 36 | 31 | 51 | 53 | 42.8 | 10.9 | 26% |
| 26 | 20 | 23 | 17 | 21.5 | 3.9 | 18% |
| 40 | 27 | 35 | 24 | 31.5 | 7.3 | 23% |
| 37 | 24 | 38 | 22 | 30.3 | 8.4 | 28% |
| 324 | 296 | 462 | 465 | 387 | 89.4 | 23% |
| 42 | 46 | 62 | 69 | 54.8 | 12.8 | 23% |
| 10 | 11 | 13 | 10 | 11.0 | 1.4 | 13% |
| 22 | 20 | 38 | 21 | 25.3 | 8.5 | 34% |
| **1439** | **1136** | **1629** | **1649** | | | |

# Citation counts for the 7 post-2000 computer science papers

| WoS | Scopus | Google Scholar | Microsoft Academic | Mean | SD | CV |
|---|---|---|---|---|---|---|
| 5 | 9 | 19 | 7 | 10.0 | 6.2 | 62% |
| 49 | 96 | 185 | 188 | 130 | 68.6 | 53% |
| 5 | 6 | 16 | 10 | 9.3 | 5.0 | 54% |
| 36 | 72 | 109 | 111 | 82.0 | 35.5 | 43% |
| 11 | 20 | 35 | 19 | 21.3 | 10.0 | 47% |
| 6 | 18 | 31 | 14 | 17.3 | 10.4 | 61% |
| 7 | 11 | 24 | 37 | 19.9 | 13.6 | 69% |
| **119** | **232** | **419** | **386** | | | |

- Web of Science, as is well known, is particularly useless when it comes to citations counts for computer science publications, since it is poor on conference proceedings where many of the most significant CS papers appear

# WHY do citation counts differ?

- For example, for our very recent conference paper
  - Silvio Peroni, David Shotton, Fabio Vitali (2016). Freedom for bibliographic references: OpenCitations arise. Proceedings of 2016 International Workshop on Linked Data for Information Extraction: 32-43.

there were 3 unique citations: One found by Scopus and all three found by Google Scholar, BUT one citing paper was duplicated in GS, with preprint and VoR having separate entries, giving it an erroneous citation count of 4

# WHY do citation counts differ?

- For example, for our very recent conference paper
  - Silvio Peroni, David Shotton, Fabio Vitali (2016). Freedom for bibliographic references: OpenCitations arise. Proceedings of 2016 International Workshop on Linked Data for Information Extraction: 32-43.

  there were 3 unique citations: One found by Scopus and all three found by Google Scholar, BUT one citing paper duplicated in GS, with preprint and VoR having separate entries, giving it an erroneous citation count of 4

- For our journal article describing DoCO, the Document Ontology
  - Constantin, A., Peroni, S., Pettifer, S., Shotton, D. and Vitali, F. (2016). The Document Components Ontology (DoCO), Semantic Web, 7: 167-181.

  there were 26 unique citations (WoS 7, Sc 11, GS 25, MA 19, D 16, SS 16). GS listed citations in six international conference papers not found by other indexes

- BUT one citing paper was duplicated in GS, with preprint and VoR having separate entries, giving a GS an apparent citation count of 25 rather than 24
- AND another citing paper was duplicated in SS, both entries with incorrect titles, giving a SS an apparent citation count of 16 rather than 15

- These varying citation counts are mostly due to differences in coverage of the literature in the four indexes, and some are due to errors


Never trust citation counts and metrics based upon them!

They are only as good as the coverage of the citation index, and are at best *relative* indicators

# Cautionary tale #2: Crossref references are incomplete

- While we are delighted that Crossref now hosts more than half a billion open journal article references, this is not the full story
- I have analysed the fate of the references in our paper

> Silvio Peroni, Alexander Dutton, Tanya Gray, David Shotton, (2015) "Setting our bibliographic references free: towards open citation data", *Journal of Documentation*, Vol. **71** Issue: 2, pp.253-277 https://doi.org/10.1108/JD-12-2013-0166
>
> Publisher: EmeraldInsight Version of Record URL: https://www.emeraldinsight.com/doi/full/10.1108/JD-12-2013-0166

# Cautionary tale #2: Crossref references are incomplete

- While we are delighted that Crossref now hosts more than half a billion open journal article references, this is not the full story
- I have analysed the fate of the references in our paper

  Silvio Peroni, Alexander Dutton, Tanya Gray, David Shotton, (2015) "Setting our bibliographic references free: towards open citation data", *Journal of Documentation*, Vol. **71** Issue: 2, pp.253-277 https://doi.org/10.1108/JD-12-2013-0166

  Publisher: EmeraldInsight Version of Record URL: https://www.emeraldinsight.com/doi/full/10.1108/JD-12-2013-0166

- Significant differences are present
  - between the authors' final manuscript and the published paper, and
  - between the published paper and the Crossref reference list
- 31 of the references in the original paper do not appear in any form in the Crossref metadata for this publication, casting into doubt the validity of any analyses based solely on open Crossref references

# The author's references

- Authors' final manuscript has 71 references, of which 43 had actionable DOIs

- Authors' benefits:
  - Provided links to 68 references: 43 hyperlinked DOIs, and 25 URIs

- Author errors and omissions:
  - Failed to find and include NCBI URL for Ref 1 (this paper has no DOI)
  - Failed to find and include DOI for Ref 67 (DOI later added by Crossref)
  - Included JSTOR URL instead of DOI for Ref 65 (DOI added by publisher)
  - Omitted in-text reference pointer for one cited paper (Vision, 2010)
  - Two errors in alphabetization of reference list (1&2; 51,52&53)

- Note: Ref 28 has no DOI or URL

# The publisher's references

- Publisher's Version of Record has 70 cited references and one "further reading", of which 32 have actionable DOIs in the form of [Crossref] links, 1 has a truncated and non-functional DOI, 7 have non-actionable DOIs included in the reference as plain text, and 13 have URLs.

- Publisher's improvements
  - Added DOI in place of author's URL for Ref 65
- Publisher's errors and omissions
  - Failed to convert 7 hyperlinked DOIs to actionable [Crossref] links, giving plain text DOIs in reference list instead (Refs 6, 19, 20, 26, 36, 49, 50)
  - Omitted DOI entirely for Ref 12
  - Truncated DOI of Ref 4, so that it does not resolve
  - For two of publisher's own papers (Refs 21 and 41) changed DOIs to URLs
  - Failed to find and include DOI for Ref 67 (later found by Crossref)
  - In updating Ref 56 from preprint to published paper, incorrectly retained "To appear in" as part of the journal title.

# The Crossref references

- Crossref has 40 references, 35 with correct DOIs, one with the incorrect truncated DOI, and 4 “unstructured” reference texts lacking a DOI
- Crossref improvements
  - Added DOI for Ref 14 (author and publisher has only URL)
  - Added DOI for Ref 67 (author and publisher has no DOI or URL)
  - Changed publisher’s URLs back to DOIs for Refs 21 and 41
- Crossref errors
  - Failed to check and correct the truncated DOI of Ref 4
  - Failed to include the DOI supplied by the publisher for Ref 59
    - Crossref lists this reference as “unstructured”, without a DOI
- The 9 other references for which the publisher failed to include actionable DOIs are completely lost from Crossref metadata (Refs 6, 19, 20, 26, 49 and 50) or are included as “unstructured” without DOIs (Refs 12, 36 and 51 )
- The 25 references to works that lack DOIs are also lost from Crossref metadata (Refs 1, 3, 8, 10, 13, 15, 16, 18, 23, 24, 27, 28, 29, 35, 37, 38, 39, 44, 46, 47, 55, 58, 66, 68, 69)

# Summary

| | Total number of references | References with actionable DOIs |
|---|---|---|
| Authors' manuscript | 71 | 43 |
| Publisher's Version of Record | 71 | 32 |
| Crossref metadata | 40 | 35 |

- So don't ever think for an instant that the Crossref open references are complete, or accurately reflect the references in the published Version of Record
  - For a start, all original references that lack DOIs (e.g. official reports and W3C Recommendations) will be missing



(To all you hackers!)

Never *never* trust the outputs of your smart algorithms without checking the source and intermediate data manually from start to finish to see where data are being lost or errors are being introduced

# An introduction to OpenCitations

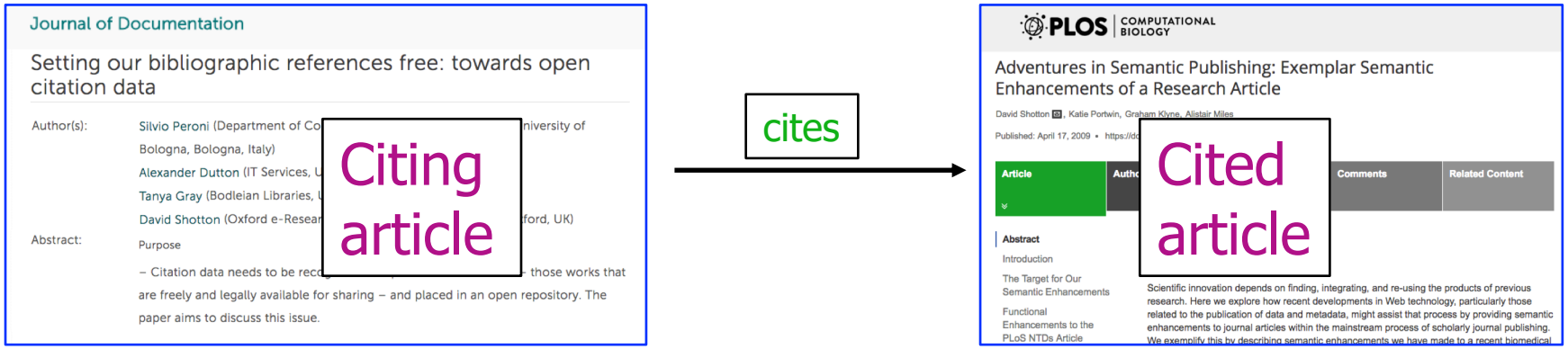
- OpenCitations (http://opencitations.net) is a scholarly infrastructure organization directed by myself and Silvio Peroni
- Our initial purpose was two-fold
  - To develop the SPAR (Semantic Publishing and Referencing) Ontologies, permitting all aspects of scholarly publishing to be described in RDF
  - To host and develop the OpenCitations Corpus (OCC), a Linked Open Data repository of bibliographic and citation data

# An introduction to OpenCitations

- OpenCitations (http://opencitations.net) is a scholarly infrastructure organization directed by myself and Silvio Peroni
- Our initial purpose was two-fold
  - To develop the SPAR (Semantic Publishing and Referencing) Ontologies, permitting all aspects of scholarly publishing to be described in RDF
  - To host and develop the OpenCitations Corpus (OCC), a Linked Open Data repository of bibliographic and citation data
- Recently, we have expanded our activities
  - We campaign for open citations as a founder member of I4OC
  - We have developed the OpenCitations data model whereby metadata for bibliographic entities and citations may be structured and organized in RDF
  - We have developed new open software of generic applicability for searching, browsing and providing REST APIs over RDF triple stores
  - We are promoting the concept of citations as first-class data entities
  - We have published a formal definition of an open citation
  - We are developing Open Citation Indexes over open citation sources

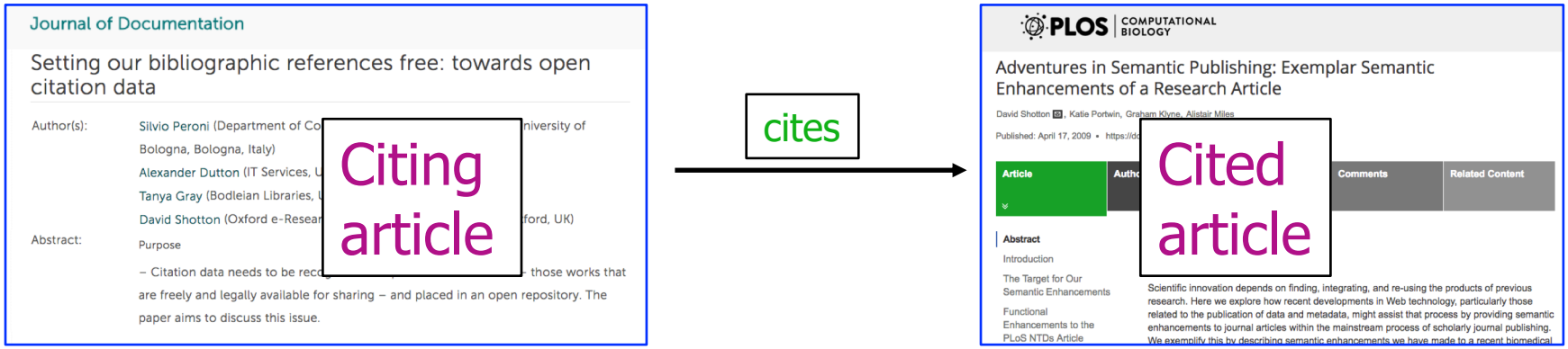
# Two views of a citation

- A citation as a simple link



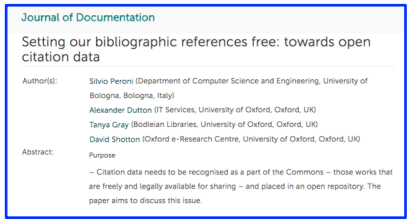
# Two views of a citation

- A citation as a simple link




- A citation as a first-class data entity

has citing article




The Citation


has cited article

with properties:

| | |
|---|---|
| has creation date | 2015 |
| has time span: | 6 years |
| has type: | Self-citation |
| has identifier: | oci:7295288-3962641 |

# How have we promoted the citation to be a first-class data entity?

1. We have published a human-readable definition of an open citation on Figshare
   - https://doi.org/10.6084/m9.figshare.6683855
2. We have made the citation definable in a machine-readable manner
   - as a member of the class “cito:Citation” with appropriate object properties
3. We have a repository for citation data, the OpenCitations Corpus
4. We have created a new global persistent identifier scheme for citations
   - The Open Citation Identifier (OCI), to parallel the DOI for publications
5. We have developed a Web-based OCI Resolution Service
   - that takes the identifier as input and returns a description of the citation

See our blog post: https://opencitations.wordpress.com/2018/02/19/citations-as-first-class-data-entities-introduction/

# Definition of an open citation

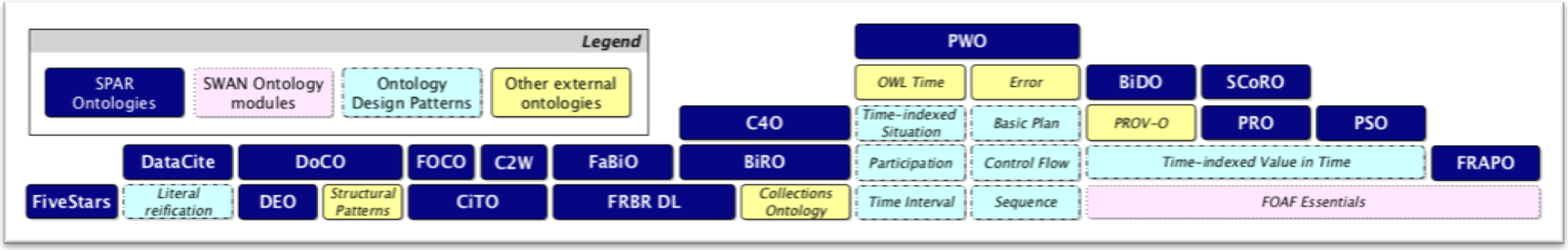
- A bibliographic citation is an open citation when the data needed to define the citation are
  - freely available, downloadable and reusable
- Specifically, such data MUST be compliant with the ‘SSO Principles’ introduced by the Initiative for Open Citations, namely that such data MUST be
  - Structured – expressed in one or more machine-readable formats
  - Separate – available without the need to access the source bibliographic entity (e.g. the article or book) in which the citation is defined
  - Open – freely accessible and reusable without restrictions

# Definition of an open citation

- A bibliographic citation is an open citation when the data needed to define the citation are
  - freely available, downloadable and reusable
- Specifically, such data MUST be compliant with the 'SSO Principles' introduced by the Initiative for Open Citations, namely that such data MUST be
  - Structured – expressed in one or more machine-readable formats
  - Separate – available without the need to access the source bibliographic entity (e.g. the article or book) in which the citation is defined
  - Open – freely accessible and reusable without restrictions
- Additionally two further principles MUST hold for the citing and cited entities
- If the citation is to be an open citation, the citing and cited entities must be
  - Identifiable using a specific persistent identifier (e.g. a DOI) or a URL
  - Available – it MUST be possible by resolving their identifiers to obtain the basic metadata of both the entities, sufficient to create or retrieve textual bibliographic references for each of them

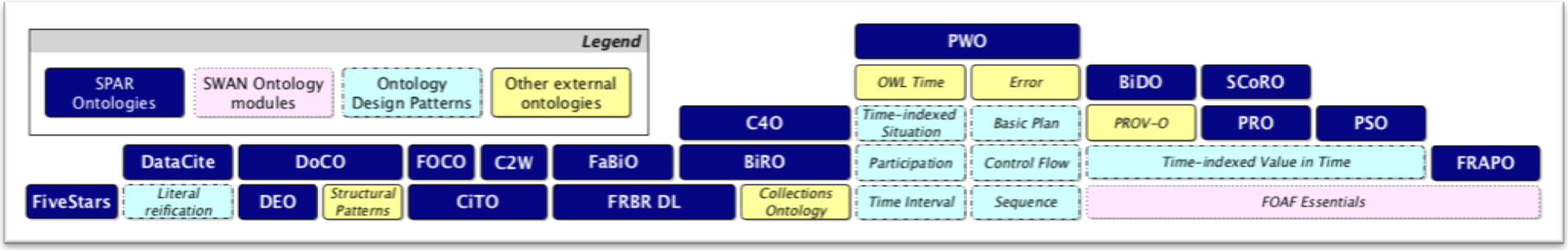
# The SPAR (Semantic Publishing and Referencing) Ontologies

- The SPAR (Semantic Publishing and Referencing) Ontologies form a suite of orthogonal and complementary OWL 2 DL ontology modules for the creation of comprehensive metadata for every aspect of semantic publishing and referencing
  - They permit metadata about scholarly artefacts to be published in machine-readable RDF, the language of interoperable Web data
  - Metadata encoded by SPAR becomes part of the Linked Open Data Web



# The SPAR (Semantic Publishing and Referencing) Ontologies

- The SPAR (Semantic Publishing and Referencing) Ontologies form a suite of orthogonal and complementary OWL 2 DL ontology modules for the creation of comprehensive metadata for every aspect of semantic publishing and referencing
  - They permit metadata about scholarly artefacts to be published in machine-readable RDF, the language of interoperable Web data
  - Metadata encoded by SPAR becomes part of the Linked Open Data Web




- Of these ontologies, CiTO and BiRO are most relevant to citations
- Our new paper describing the SPAR Ontologies suite is available at https://w3id.org/spar/article/spar-iswc2018/

# Ontology changes to permit definition of a citation in RDF

- To permit citations to be described as first-class data entities, we have made the following additions to CiTO, the Citation Typing Ontology (http://purl.org/spar/cito):
  - Class: Citation IRI: http://purl.org/spar/cito/Citation
    - Subclass: Self citation IRI: http://purl.org/spar/cito/SelfCitation

      itself with several sub-classes, e.g. Journal self citation, Author self citation
  - Data properties: has creation date and has citation time span
    - The temporal characteristic of a citation, namely the publication date of the citation (taken as the same as the publication date of the citing entity), and the date interval between that and the publication date of the cited entity
    - IRI: http://purl.org/spar/cito/hasCitationCreationDate
    - IRI: http://purl.org/spar/cito/hasCitationTimeSpan
- And we have made an addition to the DataCite Ontology (http://purl.org/spar/datacite)
  - A new member of the class Resource Identifier Scheme
    - Open Citation Identifier IRI: http://purl.org/spar/datacite/oci

# The OpenCitations Corpus: a database for citations

- The OpenCitations Corpus (OCC) is a repository of bibliographic citation data stored as Linked Open Data, with SPARQL and REST API interfaces
- The first OCC prototype was created in Oxford in 2011, with Alex Dutton as the lead developer
  - Peroni S, Dutton A, Gray T and Shotton D (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation 71:253-277. https://doi.org/10.1108/jd-12-2013-0166

# The OpenCitations Corpus: a database for citations

- The OpenCitations Corpus (OCC) is a repository of bibliographic citation data stored as Linked Open Data, with SPARQL and REST API interfaces
- The first OCC prototype was created in Oxford in 2011, with Alex Dutton as the lead developer
  - Peroni S, Dutton A, Gray T and Shotton D (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation 71:253-277. https://doi.org/10.1108/jd-12-2013-0166
- A new instance of the OCC, based on our revised OpenCitations Metadata Model, was then set up by Silvio at the University of Bologna in early July 2016
- The OpenCitations Corpus, which has been populated with scholarly references from the Open Access subset of PubMed Central
  - currently holds references from over 300,000 citing bibliographic resources
  - provides ~13 million citation links to over 6.5 million cited resources
  - these data being freely available under a CC0 public domain waiver

# The OpenCitations Corpus: a database for citations

- The OpenCitations Corpus (OCC) is a repository of bibliographic citation data stored as Linked Open Data, with SPARQL and REST API interfaces
- The first OCC prototype was created in Oxford in 2011, with Alex Dutton as the lead developer
  - Peroni S, Dutton A, Gray T and Shotton D (2015). Setting our bibliographic references free: towards open citation data. Journal of Documentation 71:253-277. https://doi.org/10.1108/jd-12-2013-0166
- A new instance of the OCC, based on our revised OpenCitations Metadata Model, was then set up by Silvio at the University of Bologna in early July 2016
- The OpenCitations Corpus, which is populated with scholarly references from the OA subset of PubMed Central
  - currently holds references from ~302,000 citing bibliographic resources, and
  - provides ~13 million citation links to over 6.5 million cited resources
  - these data being freely available under a CC0 public domain waiver
- When we resume reference ingestion after the current infrastructure upgrade, we will do so from Crossref and other sources, in addition to PubMed Central, expanding our coverage to all disciplines

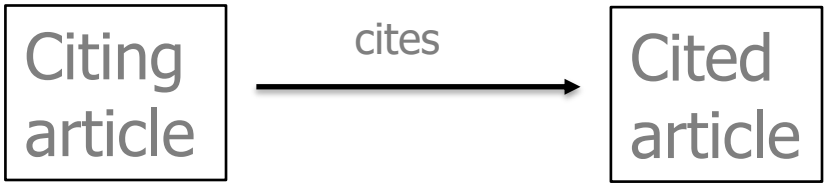
# Defining a citation from the OpenCitations Corpus in RDF

- A citation in RDF as a link


  - :citingArticle cito:cites :citedArticle .

# Defining a citation from the OpenCitations Corpus in RDF

- A citation in RDF as a link




  - :citingArticle cito:cites :citedArticle .

- A citation in RDF as an entity


  - <https://w3id.org/oc/virtual/ci/1-18> a cito:Citation ;
    cito:hasCitingEntity <https://w3id.org/oc/corpus/br/1> ;
    cito:hasCitedEntity <https://w3id.org/oc/corpus/br/18> ;
    cito:hasCitationTimeSpan "10"^^xsd:integer ;
    datacite:hasIdentifier <https://w3id.org/oc/virtual/id/ci-1-18> ;
    prov:wasAttributedTo <https://w3id.org/oc/corpus/prov/pa/7> ;
    prov:hadPrimarySource <https://w3id.org/oc/sparql?query=
    . . . [three-line query - details on application!] .

# Open Citation Identifier – the new PID for citations

- The Open Citation Identifier (OCI) is a persistent identifier for citations, paralleling the use of the DOI to identify publications

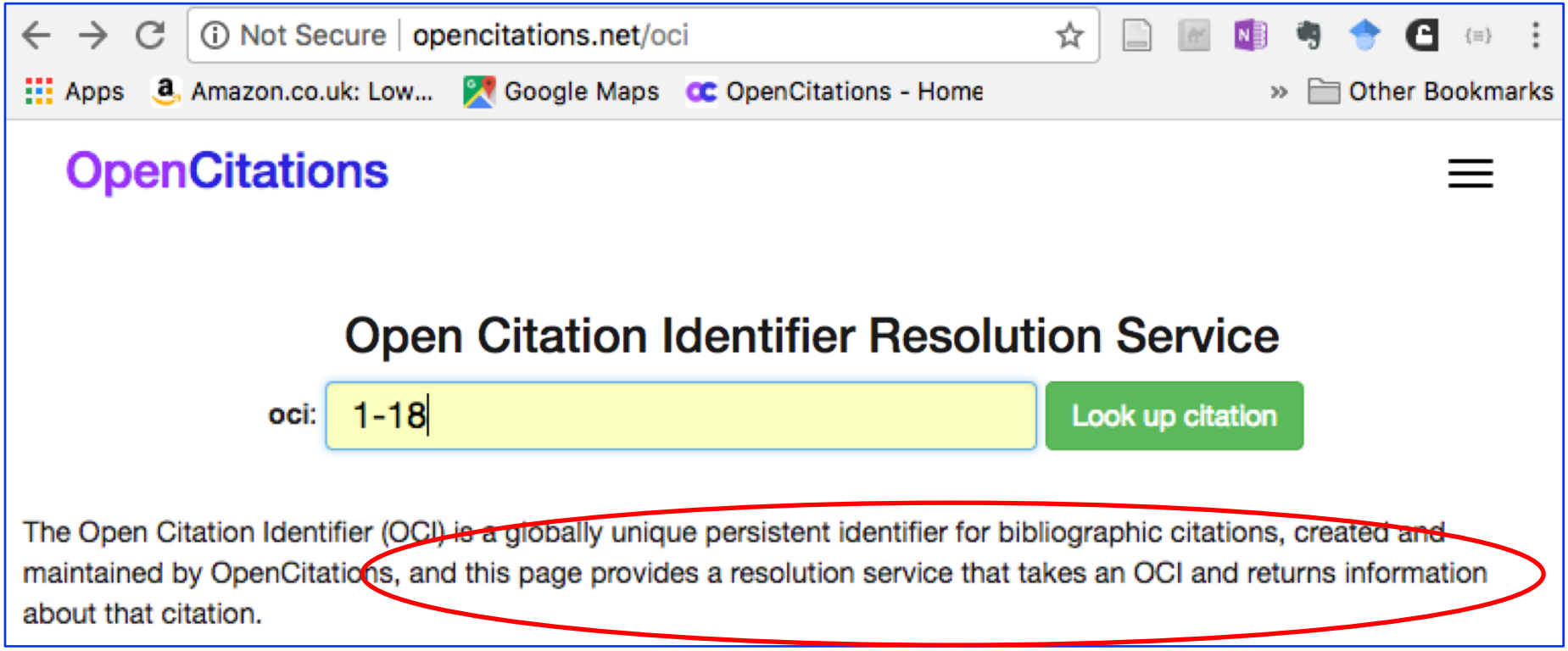
- The OCI scheme is operated by OpenCitations (http://opencitations.net/oci) and is used to identify the open citations present in the OpenCitations Corpus (OCC) *and* in other bibliographic citation databases

- Each OCI has a simple structure: oci:number-number
  - For example, oci:1-18 and oci:2544384-7295288 are both valid OCIs for citations stored in the OpenCitations Corpus
  - The first number is the OCC identifier for the citing bibliographic resource
  - The second is the OCC identifier for the cited bibliographic resource

  (These bibliographic resource identifiers are unique within the OCC)

# The Open Citation Identifier Resolution Service

- The Open Citation Identifier Resolver runs at http://opencitations.net/oci



# oci:1-18 resolves to OpenCitations metadata for the citation

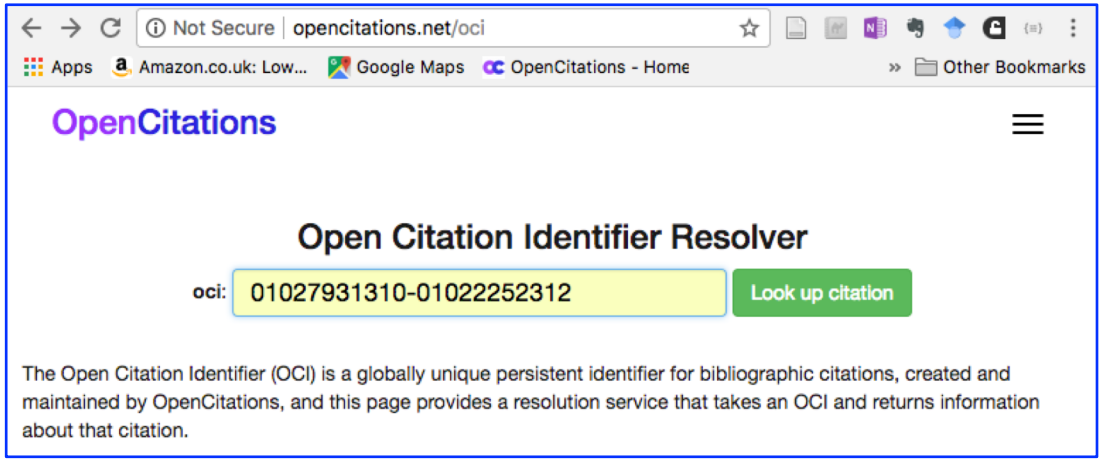

- Clicking on these links returns metadata about the citing and cited works

# The Resolver also works with external citation suppliers

- OCC bibliographic resource identifiers may contain a supplier prefix
  - a short numerical string delimited by zeros that indicates the supplier of the metadata for that bibliographic resource and its references
  - For example 010 indicates that Wikidata is the supplier of the citation data
  - Thus oci:010**27931310**-010**22252312** is the OCI for a citation between two bibliographic resources whose metadata are recorded in Wikidata
    - http://www.wikidata.org/entity/Q**27931310**, the citing resource
    - http://www.wikidata.org/entity/Q**22252312**, the cited resource

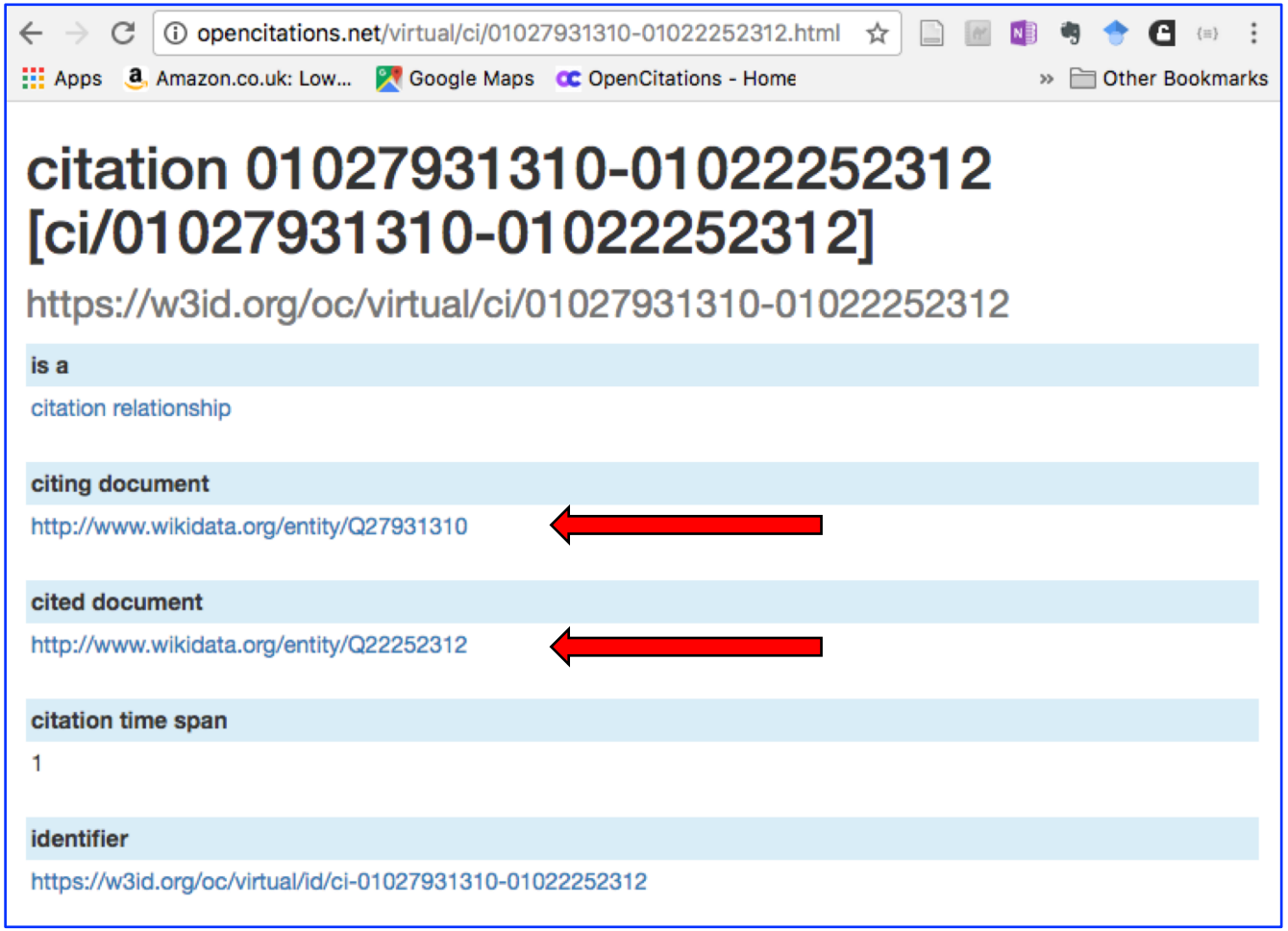
# The Resolver also works with external citation suppliers

- OCC bibliographic resource identifiers may contain a supplier prefix
  - a short numerical string delimited by zeros that indicates the supplier of the metadata for that bibliographic resource and its references
  - For example 010 indicates that Wikidata is the supplier of the citation data
  - Thus oci:01027931310-01022252312 is the OCI for a citation between two bibliographic resources whose metadata are recorded in Wikidata
    - http://www.wikidata.org/entity/Q27931310, the citing resource
    - http://www.wikidata.org/entity/Q22252312, the cited resource
- Using such an OCI in the Open Citation Identifier Resolver pulls live data from the Wikidata SPARQL endpoint and returns information about that citation



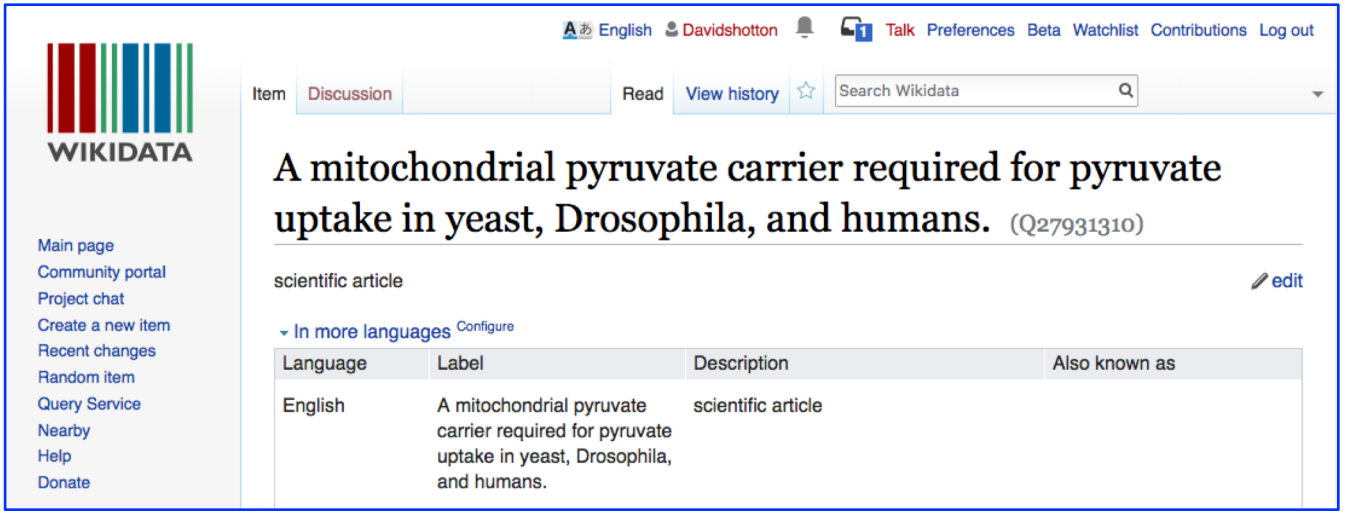
# A Wikidata citation retrieved using an Open Citation Identifier

- Resolution of oci:01027931310-01022252312

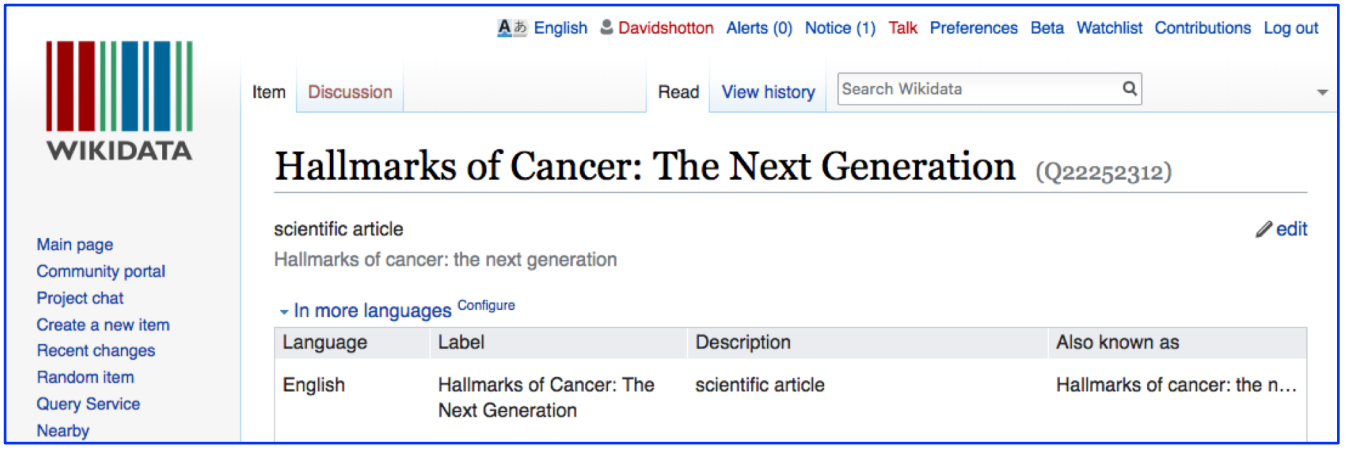


# . . . . linking to the Wikidata papers

- Citing paper




- Cited paper



# OCIs can also be used for Crossref DOI-to-DOI citations

- Initially, only numerical identifiers of bibliographic works could be used to construct the OCI

- Now they can also be constructed when works are identified using DOIs, which can contain numerals, letters and other characters
  - Each DOI is first normalized to lower case letters
  - It is then converted reversibly to a numerical string using a simple two-numeral lookup table for numerals, lower case letters and other characters (https://github.com/opencitations/oci/blob/master/lookup.csv)
    - For example, “1” becomes “01”, “2” becomes “02”, “a” becomes “10”, “b” becomes “11” and “/” becomes “36”
  - These numerical representations of DOIs are then used to create OCIs

# OCI for Crossref citations

- Thus, for the citation link in Crossref between two papers identifies by the DOIs
  - http://dx.doi.org/10.1186/1756-8722-6-59 and
  - http://dx.doi.org/10.1186/1756-8722-5-31

  the OCI is
  - oci:020010108063601070506630807020263066305 09-020010108063601070506630807020263056303 01
  - where 020 is the Crossref supplier prefix, and the green numbers are the reversible lookup table numerical equivalents of the characters in each of the DOIs, with the initial “10.” omitted

# OCI for Crossref citations

- Thus, for the citation link in Crossref between two papers identifies by the DOIs
  - http://dx.doi.org/10.1186/1756-8722-6-59 and
  - http://dx.doi.org/10.1186/1756-8722-5-31

  the OCI is
  - oci:020010108063601070506630807020263066305 09-020010108063601070506630807020263056303 01
  - where 020 is the Crossref supplier prefix, and the green numbers are the reversible lookup table numerical equivalents of the characters in each of the DOIs, with the initial "10." omitted

- In this way, OCIs have been used to create COCI, the OpenCitations Index of Crossref open DOI-to-DOI references

- In the final presentation of this Workshop, Silvio will provide technical details about COCI and other aspects of OpenCitations' work

# Acknowledgements


- Silvio Peroni

  Director of the OpenCitations Corpus,

  and the computing wizard behind everything



- The Alfred P. Sloan Foundation

which is funding our work

# Thank you for your attention! Questions???


David Shotton

Director of the OpenCitations Corpus

david.shotton@opencitations.net

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