

THE OPENCITATIONS CORPUS: AN OPEN CITATIONS DATABASE FOR SCIENTIFIC PUBLICATIONS (AS OF 2018)

Introduction

The origins of the concept of citation indexes dates back to 1955, when Eugene Garfield published in „Science” (Garfield, 1955) a proposal to create citation indexes for science as a tool of scientific journals evaluation. The first Science Citation Index (SCI) was published in 1963 and contained 102,000 articles published in 1961 in 613 selected journals. Although initially the development of bibliometric analyzes was stimulated mainly by the need to observe the development of science and the evaluation of journals, for some time quantitative bibliometric indicators have also been a principle component of parametric evaluation of scientific institutions and researchers, both in Poland and worldwide. However, the rationality of decisions determining the stimulation of scientific development that are based on quantitative indicators depends on the completeness and credibility of the sources of data from which these indicators are calculated. We can now observe the development of citation indexes from two opposite poles. The first ones are sources of a controlled quality maintained by commercial institutions working directly with publishers of scientific journals. An example of such an index is certainly SCI, founded by Eugene Garfield, which is now known at the Web of Science and is presently maintained and developed by Clarivate Analytics. Indexed data come directly from the publishers of selected scientific journals and are therefore a reliable source of metrics computed for journals or scientists. However, access is limited to paid subscribers, and the ability to download only small subsets of data for subsequent reuse (resulting from the adopted business model), make the conduct of independent bibliometric analyzes based on this source very difficult and limited. It is also worth noting that these indexes of journals are often limited to technical sciences, what makes them useless for humani-

¹ Uniwersytet Śląski w Katowicach, Instytut Bibliotekoznawstwa i Informatyki Naukowej.

ties. The second, orthogonal approach, implemented by the developers of Google Scholar, is based on acquiring declarative information directly by analyzing the content of electronic documents found on web sites of scientific institutions and publishers. As a result, the metrics presented on the Google Scholar web pages cannot be considered as definitive. However, this does not change the overall usefulness of this service in searching for specific authors or specific domains. It is worth noting that although data access is free in this case, there is also a limitation on the volume of downloadable data at one time, and the format itself doesn't help building traversable citation indexes that could enable more advanced independent bibliometric studies (Kamińska, 2017c). This indicates the need to use other methods (Kamińska, 2017d) to record bibliographic citation data, which could be used to carry out such studies (Kamińska, 2017b).

By discussing the functional aspects of existing citation indexes, it is hard to ignore recent dynamically developing ways of communicating scientific achievements such as self-publishing, publishing on the basis of free access, or taking into account other relationships binding individual documents and their authors other than just "citing" or "authorship" (e.g. the relationships used in Altmetrics (<https://www.altmetric.com/>)). In order to hold their current market position, existing citation index service providers will certainly need to revise their ontological models in the near future, according to which data are collected and made available, as well as their business operating models. Although its creators have not been motivated by commercial interests, open source software is becoming more and more popular and reliable, and commercial services are being developed over it to provide more advanced functionality or support, so sharing open bibliographic citation data could be used to calculate bibliometric coefficients, and lead to a demand for specialist services around them.

The Initiative for Open Citations

Open bibliographic citation services require open bibliographic source data. While most scholarly publishers submit reference lists to Crossref as participants in the CrossRef Cited-by service, the Crossref default is for these reference lists to be closed, requiring publishers to request that they be opened. Until recently, few have done this.

However, this situation is rapidly changing. The Initiative for Open Citations (I4OC) (Initiative for Open Citations, 2017), whose founding was spearheaded by Dario Taraborelli of the WikiMedia Foundation, is a collaboration between scholarly publishers, researchers, and other interested parties to promote the unrestricted availability of scholarly citation. Within a short space of time, I4OC has persuaded most of the major scholarly publishers to open their reference lists submitted to Crossref, so that the

proportion of all references submitted to Crossref that are now open has risen from 1% to over 45%.

The publishers who are now making available the reference lists from their scientific publications include:

- American Association for the Advancement of Science (AAAS),
- American Geophysical Union,
- American Physical Society,
- American Society for Biochemistry and Molecular Biology,
- American Society for Cell Biology,
- Association for Computing Machinery,
- BMJ,
- Cambridge University Press,
- Co-Action Publishing,
- Cold Spring Harbor Laboratory Press,
- Copernicus GmbH,
- De Gruyter,
- Edinburgh University Press,
- eLife,
- EMBO Press,
- Faculty of 1000, Ltd.,
- Frontiers Media SA,
- Geological Society of London,
- Hamad bin Khalifa University Press (HBKU Press),
- Hindawi,
- IOS Press,
- International Union of Crystallography,
- Leibniz Institute for Psychology Information,
- MDPI,
- MIT Press,
- PeerJ,
- Pensoft Publishers,
- Proceedings of the National Academy of Sciences (PNAS),
- Portland Press,
- Public Library of Science,
- Royal Society of Chemistry,
- SAGE Publishing,
- Springer Nature,
- Taylor & Francis Group,
- The Company of Biologists,
- The Rockefeller University Press,
- The Royal Society,
- Ubiquity Press, Ltd.,
- Wiley,
- World Scientific Publishing.

OpenCitations

One aspect of the growing movement towards open scholarship has been the creation of an open database of scholarly citations, the OpenCitations Corpus (OCC), available on the Internet under URL: <http://opencitations.net> (Peroni, Shotton, Vitali, 2017). OpenCitations is quite distinct from the Initiative for Open Citations, being an infrastructure organization dedicated to making bibliographic citation data available within the OCC as Linked Open Data, encoded in RDF (JSON-LD).

The corpus was formally launched in 2010, with the vision of having a global reach that would change the landscape of publishing and scientific communication, because it aimed at providing free citation information in a way that would allow the citation network paths to be traversed by URIs.

In 2015 the OCC was updated, with a new data model, new software, and new automatic data acquisition capabilities. Currently the citation data are harvested from the Open Access Subset of PubMed Central using the Europe PubMed Central (Europe PMC, 2017) API, and are processed and stored at the Department of Computer Science and Engineering of the University of Bologna.

As of 2nd June 2018, the OpenCitations Corpus has ingested the references from 302,758 citing bibliographic resources, and contains information describing 12,830,347 citation links to 6,549,665 cited resources. Once OpenCitations has ingested all the references from the 1.7 million articles in the Open Access Subset of PubMed Central, it will then start harvesting the references from the ~16 million articles already made open at Crossref in response to the Initiative for Open Citations, and the additional articles that I4OC now encourages other publishers to open, before moving on to other sources.

The OpenCitations data model

Even the most complete database will only be a useless collection of information unless its data are stored in a well-documented machine-readable format that makes possible their proper use. Various models for the organization of bibliographic data have been developed in the past. The OpenCitations developers have decided not to build closed solutions but follow open standards that can serve as reference models. To enable their data model to be reused by others, the new OpenCitations Data Model has been properly documented and openly published (Peroni, Shotton, 2017). This data model uses and integrates several different ontologies describing different aspects of the scholarly publishing domain. The diagram in Figure 1 illustrated this.

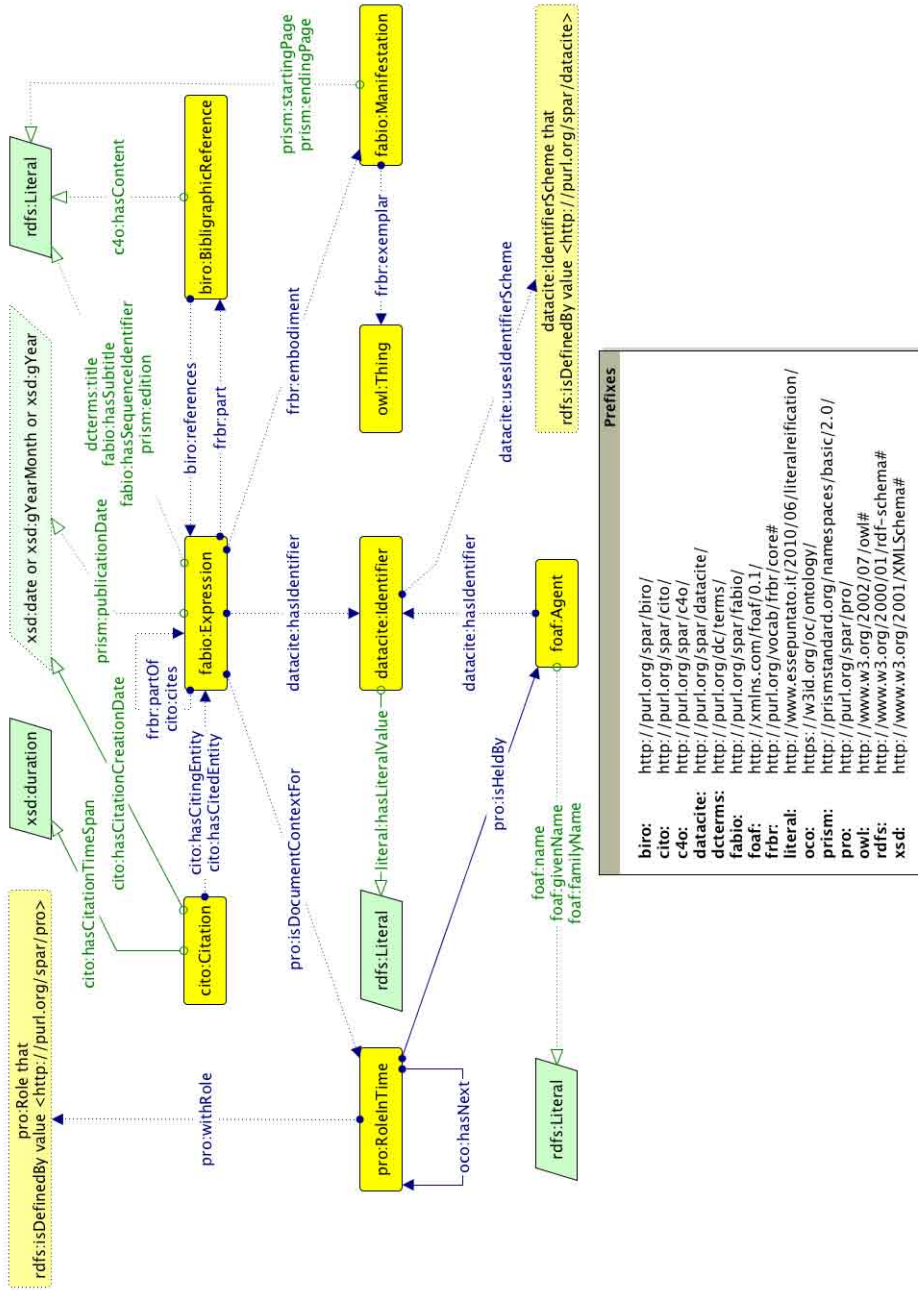


Fig. 1. A diagram of the main ontological entities described by the OCC metadata model. Source: <http://opencitations.net/model>

This diagram is made using the *Grafoo* (Graffo, 2017) tool that has been specifically created to enable graphical OWL (Web Ontology Language) modeling. This diagram, as well as the UML class diagram discussed in e.g. another author's work (Kamińska, 2018), presents a structural perspective showing how each particular domain is related. The whole concept is based on the use of a number of the SPAR (Semantic Publishing and Referencing) ontologies, specifically:

- *FaBiO (the FRBR-aligned Bibliographic Ontology)* – used to describe all the metadata of cited and citing sources and the forms in which they are included,
- *PRO (the Publishing Roles Ontology)* – used to describe the roles of agents operating on bibliographic resources (e.g. authors, publishers, editors),
- *BiRO and C4O (the Bibliographic Reference Ontology and the Citation Counting and Context Characterization Ontology)* – used to describe each of the references in the bibliography of the citation unit,
- *The Datacite Ontology* – used to describe all identifiers (e.g. DOI, ISSN, ...) of bibliographic entities and their agents.

The above information model is, for convenience, brought together into OCO (OCO, 2016), the OpenCitations Ontology, which provides integrated information about bibliographic units (both cited and citing), reference lists, agent responsibilities (e.g. author, publisher), and identifiers.

OpenCitations Functionality

While the OpenCitations platform presently lacks user-friendly graphical interfaces to visualize the citation networks it contains, it is worth noting the technology stack used to build it. All software components are freely available under the ISC or GPLv2 license, and the heart of the system is the advanced and highly scalable graph database management system *Blazegraph* (Blazegraph, 2017). It supports the *RDF* (RDF, 2017) standard (*Resource Description Framework*) which is the language for describing graph structure data, with XML-based syntax, and developed by the W3C consortium. We can obtain data from the system in several ways:

- by browsing using a simple Web interface that shows only the data about individual bibliographic entities and their references (if known) (e.g. <https://w3id.org/oc/corpus/br/1> or <http://opencitations.net/browser/br/1> for more user friendly layout),
- by downloading from Figshare data dumps of the entire corpus or corpus sub-datasets, together with their provenance data (data are updated in monthly cycles) as ZIP archives, containing JSON records for individual data areas (i.e. for persons, their roles (e.g. author, editor, publisher), individual bibliographic entities (e.g. journal articles), bibliographic

entity containers (e.g. journals), bibliographic references, identifiers (e.g. DOI, ORCID, PubMedID), and publishing forms,

- by downloading the entire corpus as an archive containing both the data and software, enabling the entire system to run on a separate hardware infrastructure,

- by running SPARQL queries (<http://opencitations.net/sparql>) to download only the subset of data (possible formats: RDF, Table, PivotTable, Google Chart) that meet the query criteria specified in SPARQL (SPARQL Query Language for RDF, 2017), a query language dedicated to processing network structure data written in RDF, of which a sample is shown in fig.2,

- by sending a SPARQL query directly to the server's listening process, omitting the graphical user interface,

- by searching the entire corpus through filters defined on attributes of document or author, what is shown in fig. 3.

The OpenCitations directors, aware of its current functional limitations in terms of data visualization, request that at this moment users

The screenshot shows the OpenCitations website's SPARQL endpoint GUI. At the top, there is a search bar and a navigation menu with links: Home, About, Corpus, Model, Download, Sparql (highlighted), Search, Oci, Publications, Licenses, and Contacts. The main heading is "OCC SPARQL endpoint GUI". Below this is a text area containing a SPARQL query:

```

1 PREFIX cito: <http://purl.org/spar/cito/>
2 PREFIX dcterms: <http://purl.org/dc/terms/>
3 PREFIX datacite: <http://purl.org/spar/datacite/>
4 PREFIX literal: <http://www.essepuntato.it/2010/06/literalreification/>
5 PREFIX biro: <http://purl.org/spar/biro/>
6 PREFIX frbr: <http://purl.org/vocab/frbr/core#>
7 PREFIX c4o: <http://purl.org/spar/c4o/>
8 SELECT ?cited ?cited_ref ?title ?url WHERE {
9   <https://w3id.org/oc/corpus/br/1> cito:cites ?cited .
10  OPTIONAL {
11    <https://w3id.org/oc/corpus/br/1> frbr:part ?ref .
12    ?ref biro:references ?cited ;
13    c4o:hasContent ?cited_ref
14  }

```

Below the query, there are four buttons: "Raw Response", "Table" (selected), "Pivot Table", and "Google Chart". At the bottom of the page, there are logos for the Dipartimento di Informatica - Scienza e Ingegneria and the University of Oxford, along with social media icons for email, Twitter, Facebook, and LinkedIn.

Fig. 2. Example of a SPARQL query over the OpenCitations Corpus. Source: <http://opencitations.net/sparql>

Fig. 3. Example of a search over the OpenCitations Corpus. Source: <http://opencitations.net/search>

wishing to interrogate the corpus should rely either on queries in SPARQL language or the very simple Web interface. However, they hope soon (the search functionality is actually the implementation of some of the planned extensions) to expand upon the present user interface capabilities (Shotton, 2017) using funds recently provided by the Alfred P. Sloan Foundation (Alfred P. Sloan Foundation, 2017).

Recently, the OpenCitations directors have introduced the OCI service. The Open Citation Identifier (OCI) is a globally unique persistent identifier for bibliographic citations, created and maintained by OpenCitations. Thanks to page show in fig. 4 it is possible to resolve the OCI to the information about citation it defines. The structure of the OCI consists of the “oci:” prefix, followed by the pair of bibliographic resource identifiers separated by a dash. The first number is the identifier for the citing bibliographic resource, while the second number is the identifier for the cited bibliographic resource.

Summary

Although the OpenCitations initiative dates back to 2010, it is its modern instantiation created in the past two years that make it stand out from alternative solutions. Basing the OpenCitations data model on documented

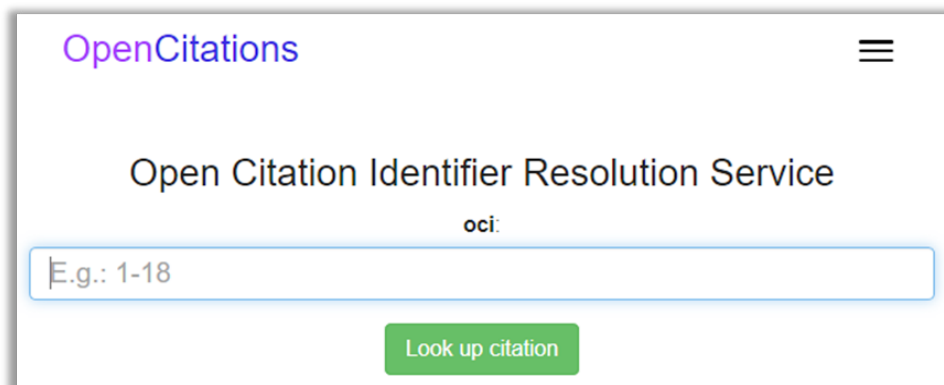
The image shows a web interface for the OpenCitations Open Citation Identifier Resolution Service. At the top left, the logo "OpenCitations" is displayed in purple. To the right of the logo is a hamburger menu icon. Below the logo, the title "Open Citation Identifier Resolution Service" is centered. Underneath the title, the label "oci:" is positioned above a text input field. The input field contains the placeholder text "E.g.: 1-18". Below the input field is a green button with the text "Look up citation".

Fig. 4. OCI: Open Citation Identifier Resolution Service. Source: <http://opencitations.net/oci>

standards makes it possible for the OpenCitations Corpus to be loaded easily with information from new sources, and for the data extracted from it to be interpreted unambiguously.

The discussion of whether the initiatives of national (Polish) bibliographic databases should be centralized or rather federated (Bednarek-Michalska, 2017) wouldn't matter if there was a common ontology model for the data stored in them. Data stored in this way and provided with global identifiers for the whole initiative can in principle be similarly used for analysis, whether they are stored in centralized or distributed form (for example, SPARQL is already capable of querying multiple sources simultaneously). Unfortunately, national initiatives also face more fundamental problems (Kamińska, 2017c) such as the need for de-duplication of duplicated bibliographic references (examples and ways of identifying and counteracting such anomalies can be found in: Kamińska, 2017a), and the inability to extract data in any structured form that allows for further analysis, since at present data are usually provided only through the user's graphical interfaces, i.e. as HTML pages.

Given the above, while it is difficult to predict when this OpenCitations initiative will prove more interesting to researchers than do commercial citation indexes or whether the graph database platform (Blazegraph) used for storing citation data will perform well (in terms of performance and stability) with increasing collected data volume, it is certainly worth getting acquainted with the general approach employed for its construction (quite apart from the specific standards and database management system employed) when planning for the modernization of similar platforms in national systems.

Bibliography

- Alfred P. Sloan Foundation. (2017). Retrieved 16 sierpnia 2017, from: <https://sloan.org/>
- Bednarek-Michalska, B. (2017). Bibliograficzne bazy danych: perspektywy i problemy rozwoju. Sprawozdanie z konferencji w Krakowie, 26–27 czerwca 2017 r. *Biuletyn EBiB*, 173. Retrieved 16 August 2017, from: <http://open.ebib.pl/ojs/index.php/ebib/article/view/544>
- Blazegraph. (2017). Retrieved 16 August 2017, from: <https://www.blazegraph.com/>
- Europe PMC. (2017). Retrieved 16 August 2017, from: <http://europepmc.org/>
- Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, 122(3159), 108-111.
- Graffo: Graphical Framework for OWL Ontologies. (2017). Retrieved 16 August 2017, from: <http://www.essepuntato.it/graffoo/>
- Initiative for Open Citations. (2017). Retrieved 16 August 2017, from: <https://i4oc.org>
- Kamińska, A.M. (2017a). Miary podobieństw łańcuchów znakowych a deduplikacja rekordów w bibliograficznych bazach danych. *Przegląd Biblioteczny*, 85(4), 477-495.
- Kamińska, A.M. (2017b). Od druków źródłowych po mapy nauki. Bibliograficzna baza danych GRUBA. W: M. Kowalska, V. Osińska (eds.), *Wizualizacja informacji w humanistyce* (pp. 17-36). Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.
- Kamińska, A.M. (2017c). ProBIT – prospektywna metoda tworzenia trawersowanych indeksów cytowań a współczesne problemy organizacji przestrzeni informacji w tradycyjnych bibliograficznych bazach danych. *Zagadnienia Informatyki Naukowej*, 55(1/109), 66-82.
- Kamińska, A.M. (2017d). Tam, gdzie zaczyna się bibliometria, czyli jak pozyskać materiał analityczny z autopsji. *Biuletyn EBiB*, 173. Retrieved 16 August 2017, from: <http://open.ebib.pl/ojs/index.php/ebib/article/view/534>
- Kamińska, A.M. (2018). O rozwoju graficznych języków komunikacji. *Zagadnienia Informatyki Naukowej*, 56(2/12) (in print)
- OCO, the OpenCitations Ontology. (2017). Retrieved 16 August 2017, from: <http://opencitations.net/ontology.html>
- Peroni, S., Shotton, D. (2017). Metadata for the OpenCitations Corpus. Retrieved 16 August 2017, from: <https://dx.doi.org/10.6084/m9.figshare.3443876>
- Peroni, S., Shotton, D., Vitali, F. (2017). One year of the OpenCitations Corpus: Releasing RDF-based scholarly citation data into the Public Domain. In: *Proceedings of the 16th International Semantic Web Conference (ISWC 2017)*. Retrieved 16 August 2017, from: <https://w3id.org/people/essepuntato/papers/oc-iswc2017.html>
- RDF. (2017). In: W3C. Retrieved 16 sierpnia 2017, from: <https://www.w3.org/RDF/>
- Shotton, D. (2017). The Sloan Foundation funds OpenCitations. In: *OpenCitations*. Retrieved 16 August 2017, from: <https://opencitations.wordpress.com/>
- SPARQL Query Language for RDF. (2017). In: W3C. Retrieved 16 August 2017, from: <https://www.w3.org/TR/rdf-sparql-query/>

Anna Małgorzata Kamińska

***The OpenCitations Corpus: an open citations database for scientific publications
(as of 2018)***

Abstract

Despite the increasingly popular sharing of scientific knowledge by open access and other “Science 2.0” concepts, scientific papers published in this way are still only “complementary” activities of researchers. This is especially true in cases where the evaluation of scientific activity is based on “strict” bibliometric indicators calculated on the basis of “recognized” data sources, which usually gather information only about works published by “prestigious” publishers. Calculation of similar metrics for publishers / platforms supporting publishing under the “Science 2.0” concepts is currently hampered by the difficulty in calculating the cumulative value of the metrics presented on various platforms. This article presents the OpenCitations Corpus, which uses recognized standards, employs the leading bibliographic ontologies, and collects and provides citation data in the form of linked open data graph structures. Open standards and free access to its collected data make the OpenCitations initiative an interesting direction for the integration of different citation data sources, allowing for cumulative / global metrics to be calculated, and the ability to download the entire corpus of data from this database will open the prospects for conducting more sophisticated bibliometric studies by interested researchers.

Key words: OpenCitations Corpus, citation index, bibliometrics, data sources, open access, bibliographic references, linked open data, RDF

Anna Małgorzata Kamińska

***OpenCitations Corpus: otwarta baza danych cytowań publikacji naukowych
(stan na 2018 r.)***

Streszczenie

Pomimo coraz bardziej popularnego dokumentowania badań naukowych i dzielenia się wiedzą na zasadach otwartego dostępu oraz upowszechniania innych koncepcji związanych z „Nauką 2.0”, publikowane w ten sposób prace naukowe stanowią jedynie „uzupełniające” działania naukowców. Dzieje się tak zwłaszcza w przypadkach, gdy ocena działalności badawczej opiera się na „twardych” wskaźnikach bibliometrycznych obliczanych na podstawie „uznanych” źródeł danych, które gromadzą informacje jedynie o pracach publikowanych przez „prestżowe” wydawnictwa. Wylizanie podobnych wskaźników dla wydawców/platform wspierających publikowanie zgodne z koncepcjami „Nauki 2.0” jest obecnie utrudnione ze względu na problemy w obliczaniu łącznej wartości będącej wypadkową wartości wskaźników prezentowanych na różnych platformach. W artykule przedstawiono inicjatywę OpenCitations Corpus, która korzysta z uznanych standardów i stosuje wiodące ontologie bibliograficzne oraz gromadzi i udostępnia dane w postaci struktur grafów cytowań. Otwarte standardy i swobodny dostęp do gromadzonych danych sprawiają, że inicjatywa OpenCitations jest interesującym kierunkiem integracji różnych źródeł danych o cytowaniach, umożliwiającym obliczanie skumulowanych/globalnych wskaźników, a możliwość pozyskania całego korpusu zgromadzonych danych otwiera perspektywy dla prowadzenia bardziej zaawansowanych badań bibliometrycznych przez zainteresowanych badaczy.

Słowa kluczowe: bibliografia załącznikowa, bibliometria, indeks cytowań, linked open data, OpenCitations Corpus, otwarty dostęp, RDF, źródła danych