Bibliotheca Nostra.
Śląski Kwartalnik Naukowy
nr 1 (51) 2018, s. 168-186

Anna Małgorzata Kamińska[1]

Plos one – a case study of quantitative and dynamic citation analysis of research papers based on the data in an open citation index (the OpenCitations Corpus)

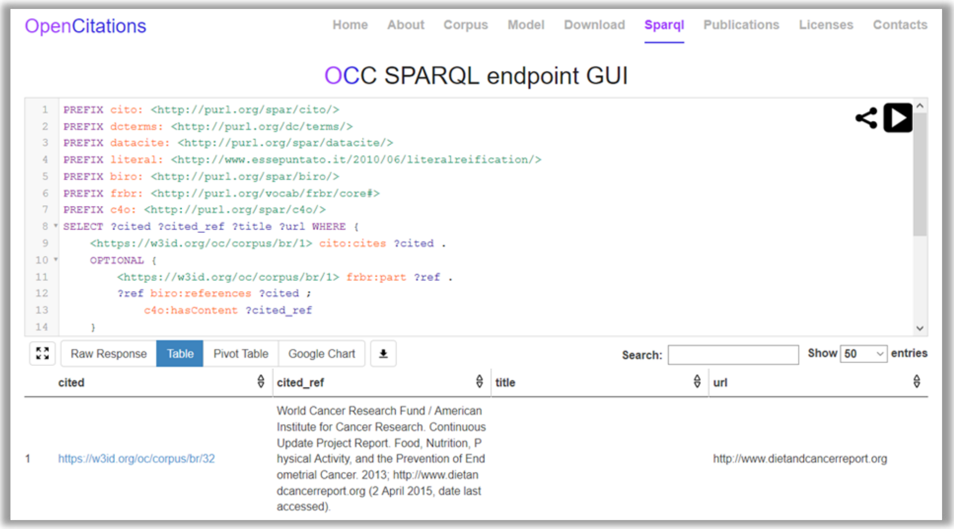
Introduction

Contemporary trends of scientific research documentation based on open access rules start changing gradually the landscape of scientometrics. The current hegemony of commercial service providers for bibliometric analyzes or bibliographic data sharing may be disturbed by the fact that the publishers more often freely distribute these data to all interested parties through dedicated web platforms. The list of such publishers is already very substantial and it seems only a matter of time when the rest of them will stop the pressures exerted by the scientific community. This situation creates the conditions for the development of non-commercial citation index systems, of which the interesting proposition seems to be OpenCitations Corpus. More about the concept itself, the model of data and the applied technologies can be found in the description by its director (Peroni, Dutton, Gray, Shotton, 2015), or in the Polish author's review paper (Kamińska, 2017a).

The purpose of this paper is to present a specific case study of the implementation of author's bibliometric analyzes based on the data collected from the OpenCitations Corpus related to selected journal of US Public Library of Science (PLOS), which together with the UK BioMed are among the largest ones in terms of publishing on open access rules.

Using the resources provided by OpenCitations Corpus is easy and intuitive thanks to a well-documented conceptual model (bibliographic ontologies) according to which the data are collected in graph database management system in a form of sentences consisting of so-called triples representing a subject, a predicate and an object (for example: <given paper><cites><another paper> or <given paper><is in><given issue of a jour-

[1] Uniwersytet Śląski w Katowicach, Instytut Bibliotekoznawstwa i Informacji Naukowej.



```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX datacite: <http://purl.org/spar/datacite/>
PREFIX literal: <http://www.essepuntato.it/2010/06/literalreification/>
PREFIX biro: <http://purl.org/spar/biro/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>
PREFIX c4o: <http://purl.org/spar/c4o/>
SELECT ?cited ?cited_ref ?title ?url WHERE {
    <https://w3id.org/oc/corpus/br/1> cito:cites ?cited .
    OPTIONAL {
        <https://w3id.org/oc/corpus/br/1> frbr:part ?ref .
        ?ref biro:references ?cited ;
            c4o:hasContent ?cited_ref
    }
```

Search: Show 50 entries

	cited	cited_ref	title	url
1	https://w3id.org/oc/corpus/br/32	World Cancer Research Fund / American Institute for Cancer Research. Continuous Update Project Report. Food, Nutrition, P hysical Activity, and the Prevention of End ometrial Cancer. 2013; http://www.dietan dcancerreport.org (2 April 2015, date last accessed).		http://www.dietandcancerreport.org

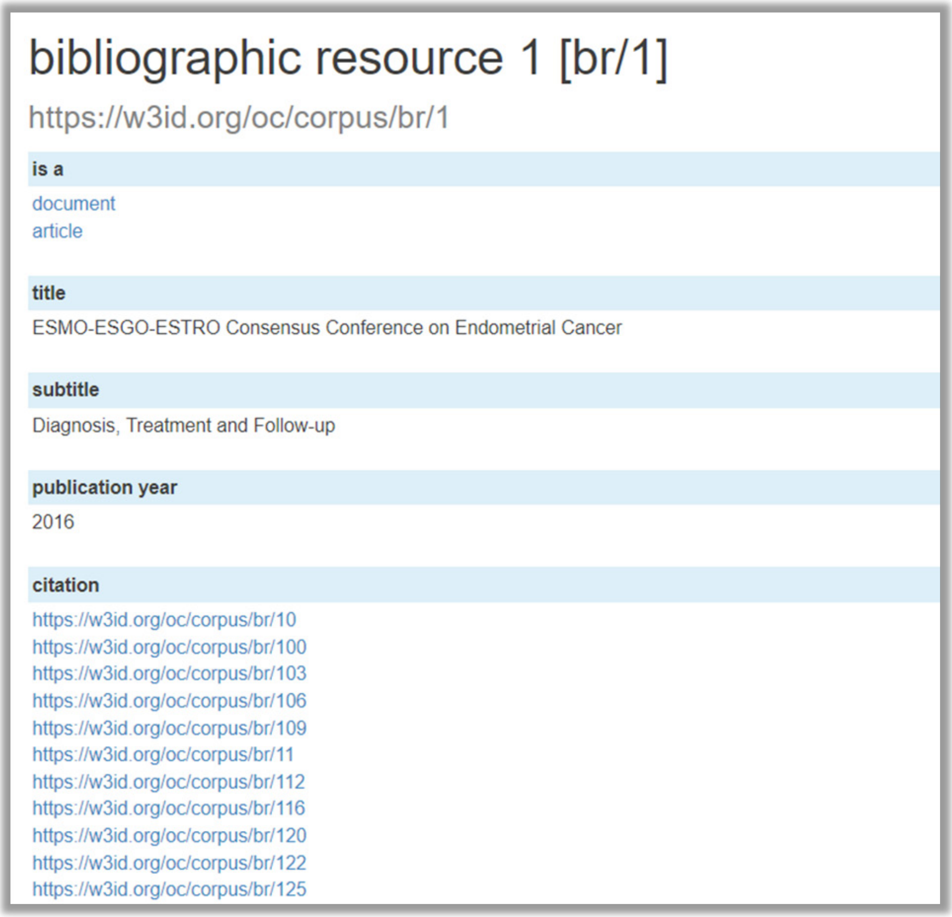
Fig. 1. The „Sparql" tab of the OpenCitations platforms. Source: http://opencitations.net/sparql

nal>). This is a technique commonly used to define semantic networks, often described in RDF (RDF, 2017), and the knowledge gathered by such description can be discovered using SPARQL (SPRQL Query Language for RDF, 2017), which allows for a variety of analytical queries.

OpenCitations Website

The basic model of data analysis is based on the use of the resources provided by the OCC website. By using the "Sparql" tab (Fig.1), one can query the OCC system and obtain the answers as an attributes list, in one of the proposed data exchange formats.

However, it is important to note that at this time, the hardware resources on which the platform services are running are modest, which may result in extended response times or hangs out of a query execution. In addition, retrieving the results of responses containing many thousands of records can be very troublesome, for example, due to an unexpected interruption of result file downloading. However, it does not change the fact that for simple queries or citations path traversing the OpenCitations platform is sufficiently complete, and identifying individual resources in accordance with the URI (Uniform Resource Identifier) concept causes navigation from the web browser through the citation paths or tracing any other relationship is easy and intuitive (Fig. 2).



bibliographic resource 1 [br/1]

https://w3id.org/oc/corpus/br/1

is a

document
article

title

ESMO-ESGO-ESTRO Consensus Conference on Endometrial Cancer

subtitle

Diagnosis, Treatment and Follow-up

publication year

2016

citation

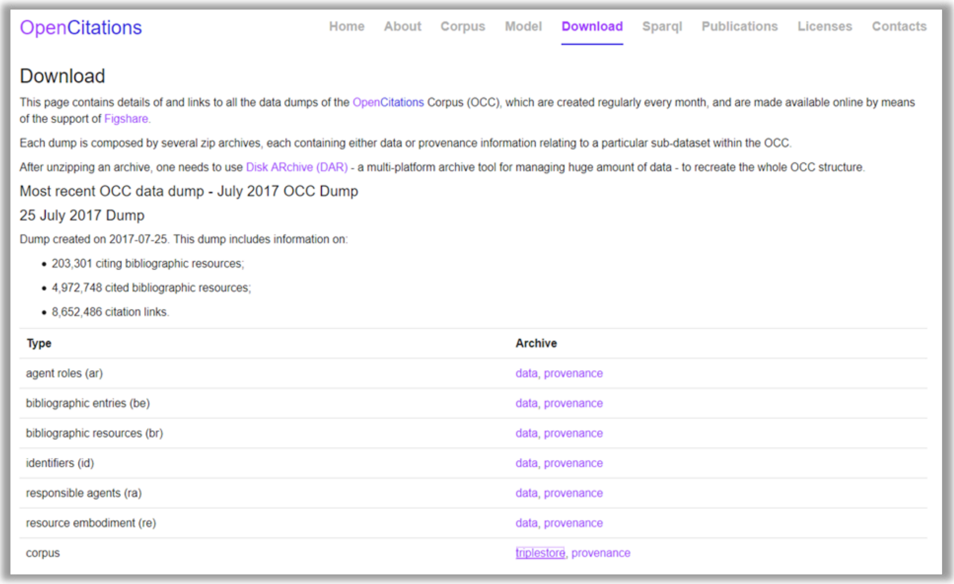
https://w3id.org/oc/corpus/br/10
https://w3id.org/oc/corpus/br/100
https://w3id.org/oc/corpus/br/103
https://w3id.org/oc/corpus/br/106
https://w3id.org/oc/corpus/br/109
https://w3id.org/oc/corpus/br/11
https://w3id.org/oc/corpus/br/112
https://w3id.org/oc/corpus/br/116
https://w3id.org/oc/corpus/br/120
https://w3id.org/oc/corpus/br/122
https://w3id.org/oc/corpus/br/125

Fig. 2. Web browser window of data preview for a given bibliographic entity. Source: http://opencitations.net/corpus/br/1.html

To download the entire corpus content, updated on a monthly basis, the “Download” tab can be used, where the “triplesto” link (Figure 3) will redirect to the appropriate “Figshare” website from which one can download the ZIP archive (with the current volume above 20GB), containing both data and software needed to run one’s own database server instance.

Configuration and deployment of BlazeGraph platform

To deploy one’s own database instance, one need an operating system platform with an installed JVM (Java Virtual Machine) environment. The



Download

This page contains details of and links to all the data dumps of the OpenCitations Corpus (OCC), which are created regularly every month, and are made available online by means of the support of Figshare.

Each dump is composed by several zip archives, each containing either data or provenance information relating to a particular sub-dataset within the OCC.

After unzipping an archive, one needs to use Disk ARchive (DAR) - a multi-platform archive tool for managing huge amount of data - to recreate the whole OCC structure.

Most recent OCC data dump - July 2017 OCC Dump

25 July 2017 Dump

Dump created on 2017-07-25. This dump includes information on:

- 203,301 citing bibliographic resources;
- 4,972,748 cited bibliographic resources;
- 8,652,486 citation links.

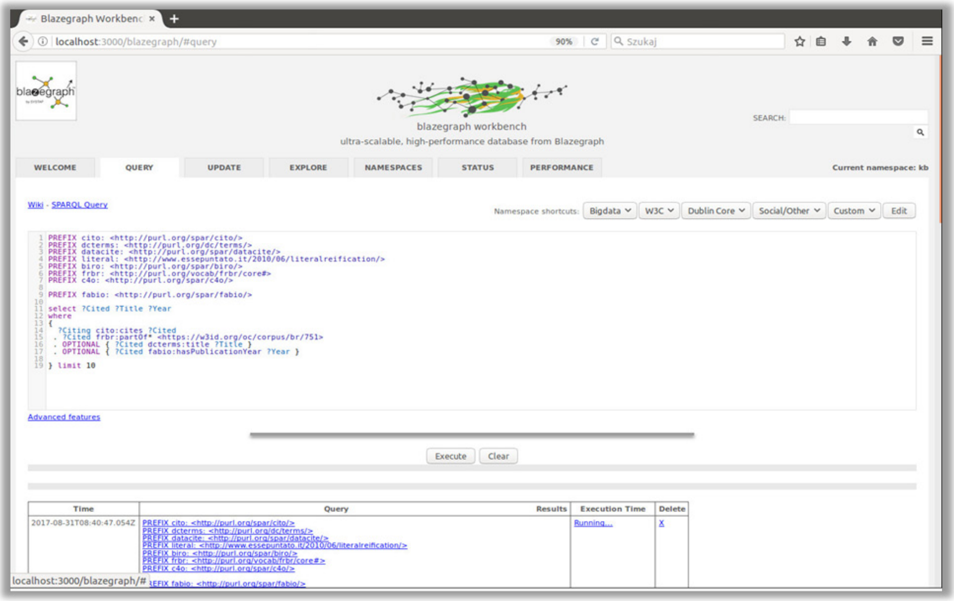
Type	Archive
agent roles (ar)	data, provenance
bibliographic entries (be)	data, provenance
bibliographic resources (br)	data, provenance
identifiers (id)	data, provenance
responsible agents (ra)	data, provenance
resource embodiment (re)	data, provenance
corpus	triplestore, provenance

Fig. 3. Web browser window of OCC data packages downloading. Source: http://opencitations.net/corpus/br/1.html

downloaded ZIP archive contains a number of DAR files that are part of another disk archiving system that allows to rebuild a full directory structure with original user permissions, file attributes and large files. After restoring the OpenCitations file structure using above mentioned program at the first level of directory tree a number of files with ".sh" extension can be find, where "run.sh" is a Linux system startup script that enables the BlazeGraph platform (Blazegraph, 2017) to run the database server used for hosting OCC data.

Communication with the system is possible through network services or through a simple web application that is made available by default on HTTP port number 3000. This can be done by typing http://localhost:3000/blazegraph/ in the address field of the web browser. However, the possibility of sending queries to the server is at http://localhost:3000/blazegraph/# query, as shown in Fig. 4.

Unfortunately, it is important to note that the ability to format or retrieve the results of a query displayed in a browser window is even more limited here than in a web application served by the OCC developers. Helpful here, however, are possibilities to query and retrieve results in CSV, XML and JSON formats using a network service interface that in the simplest case on the Linux family operating systems can be queried with the cURL command.



```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX datacite: <http://purl.org/spar/datacite/>
PREFIX literal: <http://www.essepuntato.it/2010/06/literalreification/>
PREFIX biro: <http://purl.org/spar/biro/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>
PREFIX c4o: <http://purl.org/spar/c4o/>

PREFIX fabio: <http://purl.org/spar/fabio/>

select ?Cited ?Title ?Year
where
{
  ?Citing cito:cites ?Cited
 . ?Cited frbr:partOf* <https://w3id.org/oc/corpus/br/751>
 . OPTIONAL { ?Cited dcterms:title ?Title }
 . OPTIONAL { ?Cited fabio:hasPublicationYear ?Year }

} limit 10
```

Time	Query	Results	Execution Time	Delete
2017-08-31T08:40:47.054Z	PREFIX cito: <http://purl.org/spar/cito/> PREFIX dcterms: <http://purl.org/dc/terms/> PREFIX datacite: <http://purl.org/spar/datacite/> PREFIX literal: <http://www.essepuntato.it/2010/06/literalreification/> PREFIX biro: <http://purl.org/spar/biro/> PREFIX frbr: <http://purl.org/vocab/frbr/core#> PREFIX c4o: <http://purl.org/spar/c4o/> REFIX fabio: <http://purl.org/spar/fabio/>		Running...	X

Fig. 4. The SPARQL tab of the BlazeGraph console. Source: own study

The BlazeGraph server manufacturer explains this in detail in extensive documentation illustrated in rich examples (Blazegraph – REST API, 2017).

The information presented so far gives the reader sufficient knowledge about how to implement a local computing environment, how to query it with SPARQL, how to view detailed data (both in a local and shared environment as a Web application) about bibliographic entities and other related objects, how to export query results in specified formats directly from a web application and how to run queries and export their results using the cURL command in a local environment. In the next part of the paper will be presented exemplary bibliometric analyzes of citations of articles published in the journal PLOS ONE. Of course, individual steps of described analysis can be treated as a template based on which one can conduct his own research on any other group of scientific papers.

Analyzes conducted directly on the OCC database using SPARQL

Within the OCC are mainly collected scientific papers from journals, so it is expected that more papers of this type will be among all citation units. In order to analyze the citation relation one can check how large amounts of data within individual types of cited units can be gathered in the corpus. To do this, the SPARQL query can be executed:

```
PREFIX cito: <http://purl.org/spar/cito/>

select ?types (count ( ?types ) as ? counts)
{
                ?citing cito:cites ?cited
    .           ?cited rdf:type ?types
}
group by ?types
order by desc ( ?counts )
```

The PREFIX command allows to define alias *cito* for the ontology described at <http://purl.org/spar/cito/>. Defining aliases is a good practice (increasing readability of queries), especially when the concepts of a given ontology are used in a query repeatedly. In brackets, the definition of a subset of the calculation source is included. The first triple will limit the result to all objects associated with the citation relationship (that is, it will return all entities and objects associated with cito: cites). For such pairs, the database will be searched to determine the type for all cited units previously found. These types will be aggregated and designate their counts, and then displayed in descending order of groups. The results are shown in Table 1.

It should be noted that all analyzes presented in this paper have been carried out for data released in 25.07.2017 and the results of calculations based on corpus data updated in subsequent months will certainly be different. The obtained data show that bibliographic units other than journal articles represent just over 2% of the total number of citations. The first line should be ignored because the unit can belong to several classes (multi-inheritance model) and the class indicated by the first row is not related to the publishing form. Of course, one could modify this query to return only values that are only for publishing, but it would be harder to write and longer to execute. The query execution time in present for was over 40 minutes.

In the next step you can check the number of articles stored in the entire body, grouped by individual publishers. To do this, you need to query:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>

SELECT  ?journaltitle ( count(*) as ?counts )
{
    ?citing rdf:type <http://purl.org/spar/fabio/JournalArticle>
  . ?citing frbr:partOf* ?container
  . ?container dcterms:title ?journaltitle
  . ?container rdf:type <http://purl.org/spar/fabio/Journal>
}
group by ?journaltitle
order by desc ( ?counts )
```

Table. 1. Citation counts by types

types	counts
<http://purl.org/spar/fabio/Expression>	8652350
<http://purl.org/spar/fabio/JournalArticle>	7270180
<http://purl.org/spar/fabio/BookChapter>	81829
<http://purl.org/spar/fabio/ProceedingsPaper>	27832
<http://purl.org/spar/fabio/Book>	17656
<http://purl.org/spar/fabio/ReferenceEntry>	16246
<http://purl.org/spar/fabio/DataFile>	6507
<http://purl.org/spar/fabio/ReportDocument>	2387
<http://purl.org/spar/fabio/Thesis>	741
<http://purl.org/spar/fabio/SpecificationDocument>	631
<http://purl.org/spar/fabio/Journal>	253
<http://purl.org/spar/fabio/Series>	193
<http://purl.org/spar/fabio/JournalIssue>	188
<http://purl.org/spar/fabio/ReferenceBook>	133
<http://purl.org/spar/fabio/ExpressionCollection>	51
<http://purl.org/spar/fabio/AcademicProceedings>	35
<http://purl.org/spar/fabio/BookSeries>	16

Source: own study

This query searches all units that are articles from journals. These magazines are contained is several "containers" ("JournalIssue", "Journal-Volume", "Journal") at subsequent levels of the hierarchy. The query filters subset of data only to entity that are related to the publisher for which the title is searched. Titles are then aggregated and displayed in decreasing order of group size.

The result is a very extensive list (over 26,000) of journals, so it was limited to the first ten items and presented below.

The results show that most articles in the OCC database were collected for the PLOS ONE journal.

If one want to limit the entire graph of citations to PLOS ONE only data, one should know the identifier (URI) of that publisher, which is given in the

Table. 2. Citation counts by journal

journaltitle	counts
PLOS ONE - PLoS ONE	93056
Proceedings of the National Academy of Sciences	49679
Journal of Biological Chemistry	42100
Sci. Rep. - Scientific Reports	27150
Science	21621
Nature	20928
The Journal of Immunology	13327
Nucleic Acids Research	13182
Journal of Neuroscience	12557
Phys. Rev. Lett. - Physical Review Letters	12210

Source: own study

corpus so that it can be used in subsequent queries. To do this, the following query can be executed:

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT * WHERE {
            ?citing rdf:type <http://purl.org/spar/fabio/Journal>
   .        ?citing dcterms:title ?title
   FILTER regex(?title, “^PLOS”)
}
```

The command displays a journal names starting with “PLOS”. This allows to find its identifier and its the value is: <https://w3id.org/oc/corpus/br/751>.

Using of SPARQL commands, further research is possible, such as the analysis of the number of cited papers from specific journals, articles from specific authors, the impact of individual journals, and more. However, probably not for everyone, this way will be the most intuitive and quickest to the intended destination. Therefore, if one wants to use other tools for analyzing graph structures, the export of data subset that is needed for conducting particular analyzes to the data structures that describe separately the vertices and edges of the citation graphs, as this data form accepts most analytical systems dedicated to network structure research.

Analyzes conducted in the Gephi platform

As an exemplary tool for further analysis the Gephi platform was chosen. Although it exists for more than 8 years, it is still in the development phase (so-called beta version). Thanks to its user-friendly GUI and easy extension of functionality (lots of plugins) it is eagerly chosen by many researchers.

Its native file format for networks is GEXF based on XML and using this format gives a number of advantages, as shown in one of the author's earlier work (Kamińska, 2018c), but from the BlazeGraph database is much easier (with the use of SPARQL queries only) to export data as a CSV file. For a file describing the edges, the Gephi application expects to have at least two columns named "Source" and "Target" containing the source and target node IDs. The result of the following SPARQL query saved to a file can directly be a source of information describing the citation we can supply to the Gephi application as CSV file:

```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>

SELECT
(
replace(str(?Citing),'https://w3id.org/oc/corpus/br/','') as ?Source
)
(
replace(str(?Cited),'https://w3id.org/oc/corpus/br/','') as ?Target
)
WHERE
{
     ?Citing cito:cites ?Cited
   . ?Cited frbr:partOf* <https://w3id.org/oc/corpus/br/751>
}
```

The query searches all citations and filters the results to the records for which the cited unit is included in the journal of the previously found identifier pointing to PLOS ONE. In the SELECT statement, also functions is added that removes prefixes specific to URIs, leaving only a numeric value to allow for a more transparent form of identifiers.

By saving the query result in CSV format, we automatically get a header with the required names "Source" and "Target", so that it can be imported with the "Import spreadsheet" function as an edge file to the Gephi application from the “Data laboratory” tab. By selecting “create missing nodes” the file can be loaded and the platform will automatically generate vertices based on the identifiers found in the “Source” and “Target” columns. This will allow to analyze the network structure, however, so generated vertices representing the papers will not contain any information (except identifiers that identify the OCC bibliographic entities) describing them. If one wants to be able to view in the Gephi basic article information along with additional information that will describe them (article title, release year, etc.), a file with the vertices should be created. Assuming that analysis conducting here focuses solely on articles from PLOS ONE journal, it is only need to generate a file describing only those bibliographic units. It can be done with the following command:

```
PREFIX cito: <http://purl.org/spar/cito/>
PREFIX frbr: <http://purl.org/vocab/frbr/core#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX fabio: <http://purl.org/spar/fabio/>

SELECT
 distinct
 ( replace(str(?Cited),’https://w3id.org/oc/corpus/br/’,”) as ?Id )
 ( ?Title as ?Label )
 ?Year
WHERE
{
     ?Citing cito:cites ?Cited
   . ?Cited frbr:partOf* <https://w3id.org/oc/corpus/br/751>
   . OPTIONAL { ?Cited dcterms:title ?Title }
   . OPTIONAL { ?Cited fabio:hasPublicationYear ?Year }
}
```

The above query returns all unique unit identifiers cited and published in PLOS ONE. In addition, if they are described by titles and publication years, this information will also be included in the results. It is worth noting that the result saved in CSV format will include the identifiers and titles labeled as headers “Id” and “Label” respectively. These are the headers expected by Gephi. On the other hand, a header with the name “Year” describing the column with years of publication will be an additional attribute describing the node.

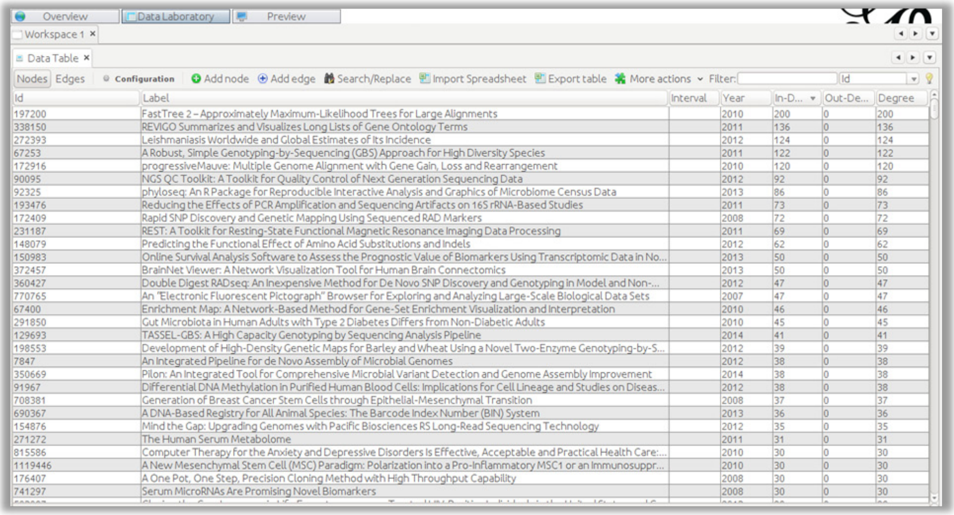
The order of steps to create a citation graph in the Gephi using the two above generated files is as follows:

1. create a new project (“New project”) in the Gephi platform,
2. import the file of cited entities (“Nodes table” option) with titles and years of publication,

3. import the file of citation edges ("Edges tabl" option) using the "Create missing node" option.

So feed up system is ready to conduct analyzing. For more information on how to import CSV data, visit the developer's website (Gephi makes graphs handy, 2017).

For the data loaded in the above described method, the number of citations for individual articles was calculated as indegree. Therefore, all citation units (both from PLOS ONE and all others) were included. As the examplary objective of the analysis is the articles published in PLOS ONE then the graph is limited to such units. The graph of citations between PLOS ONE units was obtained, but it contains information on the number of all citations per paper.



Id	Label	Interval	Year	In-D...	Out-De...	Degree
197200	FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments		2010	200	0	200
338150	REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms		2011	136	0	136
272393	Leishmaniasis Worldwide and Global Estimates of Its Incidence		2012	124	0	124
67253	A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species		2011	122	0	122
172916	progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement		2010	120	0	120
90095	NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data		2012	92	0	92
92325	phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data		2013	86	0	86
193476	Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies		2011	73	0	73
172409	Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers		2008	72	0	72
231187	REST: A Toolkit for Resting-State Functional Magnetic Resonance Imaging Data Processing		2011	69	0	69
148079	Predicting the Functional Effect of Amino Acid Substitutions and Indels		2012	62	0	62
150983	Online Survival Analysis Software to Assess the Prognostic Value of Biomarkers Using Transcriptomic Data in No...		2013	50	0	50
372457	BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics		2013	50	0	50
360427	Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-...		2012	47	0	47
770765	An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets		2007	47	0	47
67400	Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation		2010	46	0	46
291850	Gut Microbiota in Human Adults with Type 2 Diabetes Differs from Non-Diabetic Adults		2010	45	0	45
129693	TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline		2014	41	0	41
198553	Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-S...		2012	39	0	39
7847	An Integrated Pipeline for de Novo Assembly of Microbial Genomes		2012	38	0	38
350669	Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement		2014	38	0	38
91967	Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Diseas...		2012	38	0	38
708381	Generation of Breast Cancer Stem Cells through Epithelial-Mesenchymal Transition		2008	37	0	37
690367	A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System		2013	36	0	36
154876	Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology		2012	35	0	35
271272	The Human Serum Metabolome		2011	31	0	31
815586	Computer Therapy for the Anxiety and Depressive Disorders Is Effective, Acceptable and Practical Health Care:...		2010	30	0	30
1119446	A New Mesenchymal Stem Cell (MSC) Paradigm: Polarization into a Pro-Inflammatory MSC1 or an Immunosuppr...		2010	30	0	30
176407	A One Pot, One Step, Precision Cloning Method with High Throughput Capability		2008	30	0	30
741297	Serum MicroRNAs Are Promising Novel Biomarkers		2008	30	0	30

Fig. 5. The entities from PLOS ONE in descending order of citation counts. Source: own study

The citation number values are presented in Fig. 5, which shows that the most frequently cited article (200 times) is "Fast Tree ...", while the subsequent ("REVIGO Summarizes ...", "Leishmaniasis Worldwide ..." and more) have a great distance from the leader. Their number visualized as node size using a layout algorithm based on the simulation of gravity forces as the "citation map" is shown in Fig. 6.

This map allows to discover relationships hard to observe in tabular form. The edges connecting the individual nodes represent a citation relation that is not reflexive so the graph is directed and the direction of the citation is in clockwise direction of the edge. It can be seen here that the papers

"REVIGO Summarizes ..." and "Leishmaniasis Worldwide ..." are often cited, but by units not published by PLOS ONE journal. On the other hand, the related group of articles "Fast Tree ..", "A Robust ..." and others point to the possibility of their mutual thematic relationships. It is worth noting that units cited rarely have very small vertices and their labels. They represent only a less important background of analysis. While presenting a "map" as a static image, it is difficult to see the names of the smallest ones, but of course, using the Gephi tool, it is possible to interactively zoom in and focus on the selected sub-areas of the map. Based on such a "map" one can build hypotheses that can be verified by deeper look at particular areas of "map" or detailed analysis of raw data on selected bibliography entities. For the "Fast Tree ..." entity, the graph of the papers related to it is shown in Fig. 7. We see here that despite the high rankings this article is cited only by 3 other papers published in PLOS ONE (other come from different journals).

The situation is quite different in the case of the paper "Age Targeting..." for which the graph is presented in Fig. 8. Although this entity has received "only" nine citations, there is a rich network of relationship (both citing and cited links) that may point to great interest to subject matter of this paper by PLOS ONE journal.

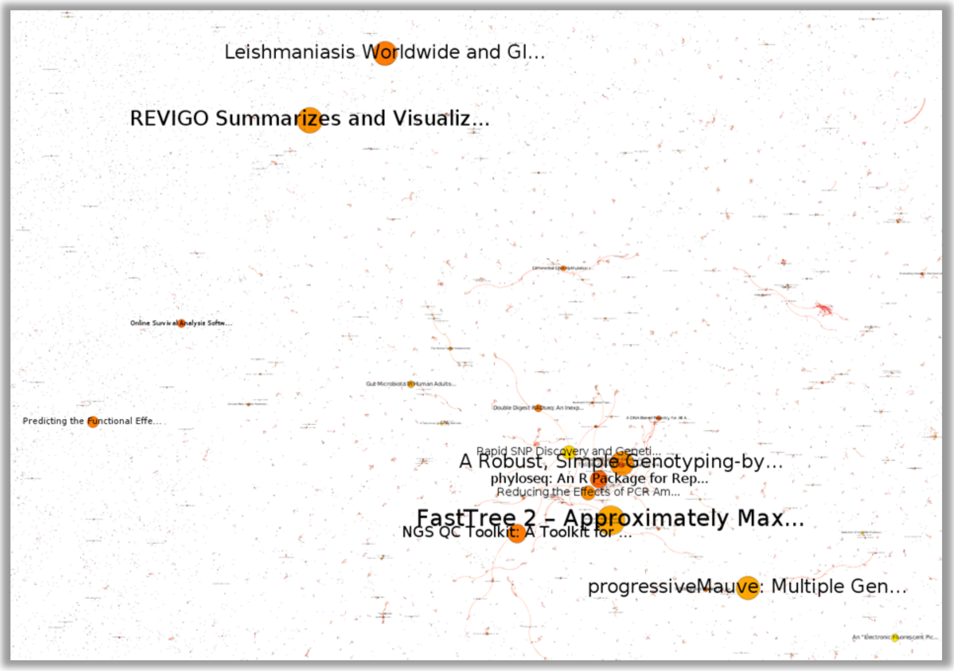

Leishmaniasis Worldwide and Gl...
REVIGO Summarizes and Visualiz...
Online Survival Analysis Softw...
Gut Microbiota in Human Adults...
Predicting the Functional Effe...
Double Digest RADseq: An Inexp...
Rapid SNP Discovery and Geneti...
A Robust, Simple Genotyping-by...
phyloseq: An R Package for Rep...
Reducing the Effects of PCR Am...
FastTree 2 – Approximately Max...
NGS QC Toolkit: A Toolkit for ...
progressiveMauve: Multiple Gen...
An "Electronic Fluorescent Pic...

Fig. 6. Citation map of papers from PLOS ONE. Source: own study

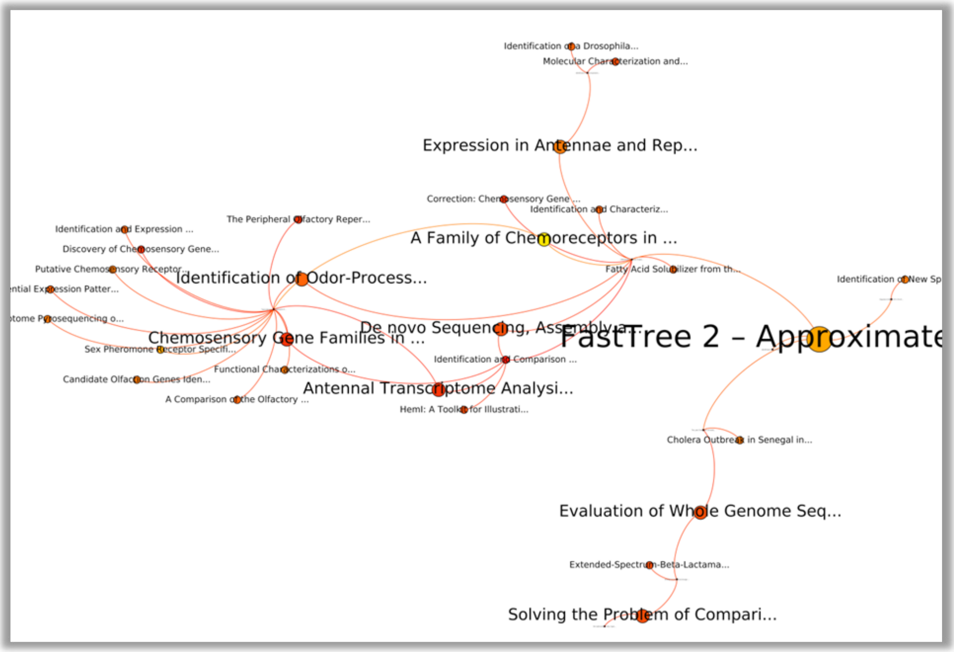

Identification of a Drosophila...
Molecular Characterization and...
Expression in Antennae and Rep...
Correction: Chemosensory Gene ...
Identification and Characteriz...
The Peripheral Olfactory Reper...
Identification and Expression ...
A Family of Chemoreceptors in ...
Discovery of Chemosensory Gene...
Fatty Acid Solubilizer from th...
Identification of New Sp
Putative Chemosensory Receptor...
Identification of Odor-Process...
ntial Expression Patter...
tome Pyrosequencing o...
De novo Sequencing, Assembly, a...
FastTree 2 – Approximate
Chemosensory Gene Families in ...
Sex Pheromone Receptor Specifi...
Identification and Comparison ...
Functional Characterizations o...
Candidate Olfaction Genes Iden...
Antennal Transcriptome Analysi...
A Comparison of the Olfactory ...
Heml: A Toolkit for Illustrati...
Cholera Outbreak in Senegal in...
Evaluation of Whole Genome Seq...
Extended-Spectrum-Beta-Lactama...
Solving the Problem of Compari...

Fig. 7. Citation map of papers related to „Fast Tree..". Source: own study

While citation maps, which represent directed and unitary citation facts between scientific papers, can represent fairly long citation paths that are sometimes difficult to visualize and analyze without the use of tools such as interactive zoom or graph filtering, the use of e.g. the co-citation measure gives maps to be easier to analyze as static images. This measure for two selected bibliographic entities assigns a value equal to the number of documents they cite these entities at the same time. The resulting graph is thus a non-directed but a weighted graph, what means that the edges connecting the nodes have no direction, but have their characteristic value indicating the strength of correlation of two given entities, that for better visualization, can affect thickness of edges. An example of visualization of this measure for entities of the analyzed corpus (a subset of PLOS ONE originating from the OCC) is illustrated in Fig. 9. It can be read in such a way that the arched documents are related to each other the more the edge is thicker.

An interesting example of visualization of dynamic in a bibliometric study is shown on the Fig. 10. The whole corpus of papers has been filtered to the most frequent cited articles. The size of nodes is proportional to citation count measure but at the same time the bigger nodes "bubbled out" at the top of the graph. The abscissa shows the partitions grouping

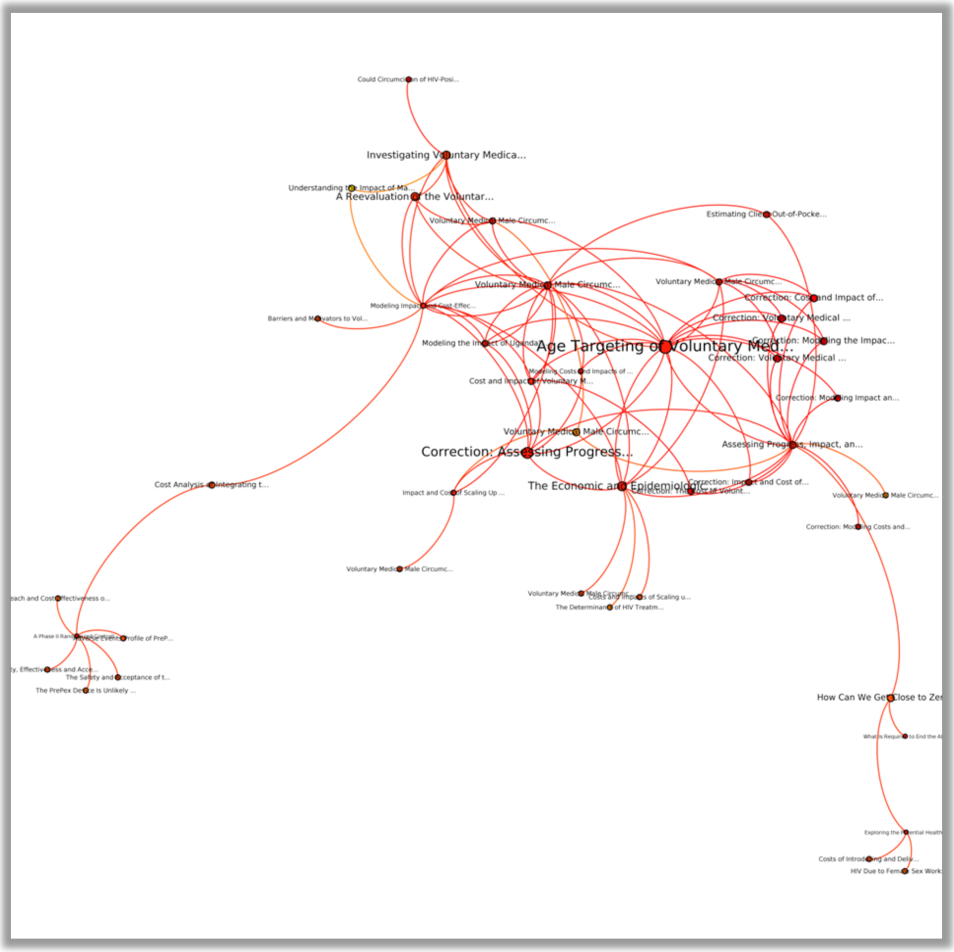

Could Circumcision of HIV-Posi...
Investigating Voluntary Medica...
Understanding the Impact of Ma...
A Reevaluation of the Voluntar...
Voluntary Medical Male Circumc...
Estimating Client Out-of-Pocke...
Voluntary Medical Male Circumc...
Voluntary Medical Male Circumc...
Correction: Cost and Impact of...
Modeling Impact and Cost-Effec...
Barriers and Motivators to Vol...
Correction: Voluntary Medical ...
Modeling the Impact of Uganda's
Age Targeting of Voluntary Med...
Correction: Modeling the Impac...
Correction: Voluntary Medical ...
Modeling Costs and Impacts of ...
Cost and Impact of Voluntary M...
Correction: Modeling Impact an...
Voluntary Medical Male Circumc...
Correction: Assessing Progress...
Assessing Progress, Impact, an...
Cost Analysis of Integrating t...
Impact and Cost of Scaling Up ...
The Economic and Epidemiologic...
Correction: Impact and Cost of...
Correction: The Cost of Volunt...
Voluntary Medical Male Circumc...
Correction: Modeling Costs and...
Voluntary Medical Male Circumc...
Voluntary Medical Male Circumc...
Costs and Impacts of Scaling u...
The Determinants of HIV Treatm...
each and Cost-Effectiveness o...
A Phase II Randomized Controlled
Adverse Events Profile of PreP...
ty, Effectiveness and Acce...
The Safety and Acceptance of t...
The PrePex Device Is Unlikely ...
How Can We Get Close to Zer
What Is Required to End the A
Exploring the Potential Health
Costs of Introducing and Deliv...
HIV Due to Female Sex Work

Fig. 8. Citation map of papers related to „Age Targeting..”. Source: own study

publications by year of issue in a range between 2006 and 2017. The arcs between nodes show citations in clockwise direction of the edge. As we can see, however, the most frequently cited articles don't have incoming arcs. This is because of the graph filtration and it proves that citation count for these papers comes from articles not published in PLOS ONE or from article published in PLOS ONE but with a relatively small citation count.

An interesting case can be observed for some articles from 2016-2017, where there are the inverted citation arcs – the articles from 2016 cites the article from 2017. In fact, it is the relationship "Erratum in" (as evidenced

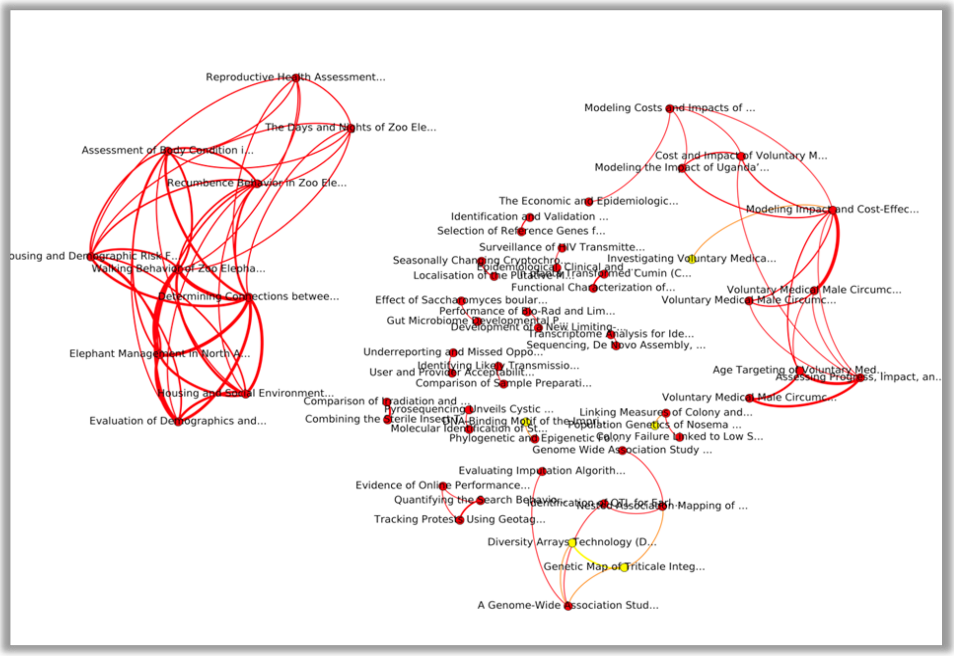

Reproductive Health Assessment...
The Days and Nights of Zoo Ele...
Assessment of Body Condition i...
Recumbence Behavior in Zoo Ele...
ousing and Demographic Risk F...
Walking Behavior of Zoo Elepha...
Determining Connections betwee...
Elephant Management in North A...
Housing and Social Environment...
Evaluation of Demographics and...
Modeling Costs and Impacts of ...
Cost and Impact of Voluntary M...
Modeling the Impact of Uganda'...
The Economic and Epidemiologic...
Identification and Validation ...
Selection of Reference Genes f...
Modeling Impact and Cost-Effec...
Surveillance of HIV Transmitte...
Seasonally Changing Cryptochro...
Epidemiological, Clinical and ...
Investigating Voluntary Medica...
Localisation of the Putative M...
In planta Transformed Cumin (C...
Functional Characterization of...
Voluntary Medical Male Circumc...
Effect of Saccharomyces boular...
Voluntary Medical Male Circumc...
Performance of Bio-Rad and Lim...
Gut Microbiome Developmental P...
Development of a New Limiting-...
Transcriptome Analysis for Ide...
Sequencing, De Novo Assembly, ...
Underreporting and Missed Oppo...
Identifying Likely Transmissio...
User and Provider Acceptabilit...
Age Targeting of Voluntary Med...
Assessing Progress, Impact, an...
Comparison of Sample Preparati...
Comparison of Irradiation and ...
Pyrosequencing Unveils Cystic ...
Voluntary Medical Male Circumc...
Combining the Sterile Insect T...
DNA-Binding Motif of the Impri...
Linking Measures of Colony and...
Molecular Identification of St...
Population Genetics of Nosema ...
Phylogenetic and Epigenetic Fo...
Colony Failure Linked to Low S...
Genome Wide Association Study ...
Evaluating Imputation Algorith...
Evidence of Online Performance...
Quantifying the Search Behavior...
Identification of QTL for Earl...
Nested Association Mapping of ...
Tracking Protests Using Geotag...
Diversity Arrays Technology (D...
Genetic Map of Triticale Integ...
A Genome-Wide Association Stud...

Fig. 9. The map of papers co-citation. Source: own study

by bibliographic data published in the source https://www.ncbi.nlm.nih.gov/pubmed/27441648), but in the OpenCitations system this has been categorized as a part of references (http://opencitations.net/browser/br/105147).

The analyzes described above done using the Gephi tool show only the basics of its use for bibliometric researches. In addition to citations between bibliographic entities, analyzes can also be made at the higher level of aggregation (e.g. journals or institutions), analysis of co-operation between researchers (both in terms of citation or co-authorship), visualization of bibliometric indicators such as number of citations, measure of bibliographic coupling or measure of co-citation. Each of these issues on the example of data from the Polish bibliographic database CYTBIN was presented in the author's earlier work (Kamińska, 2017c), and the issue of co-authorship visualization has been more extensively presented on the basis of the GRUBA bibliographic database developed by the author (Kamińska, 2018c).

It is worth mentioning that the Gephi platform also allows to calculate many of the metrics used in social network analysis (SNA). Proposals for the use of these measures on the basis of bibliometrics and on other

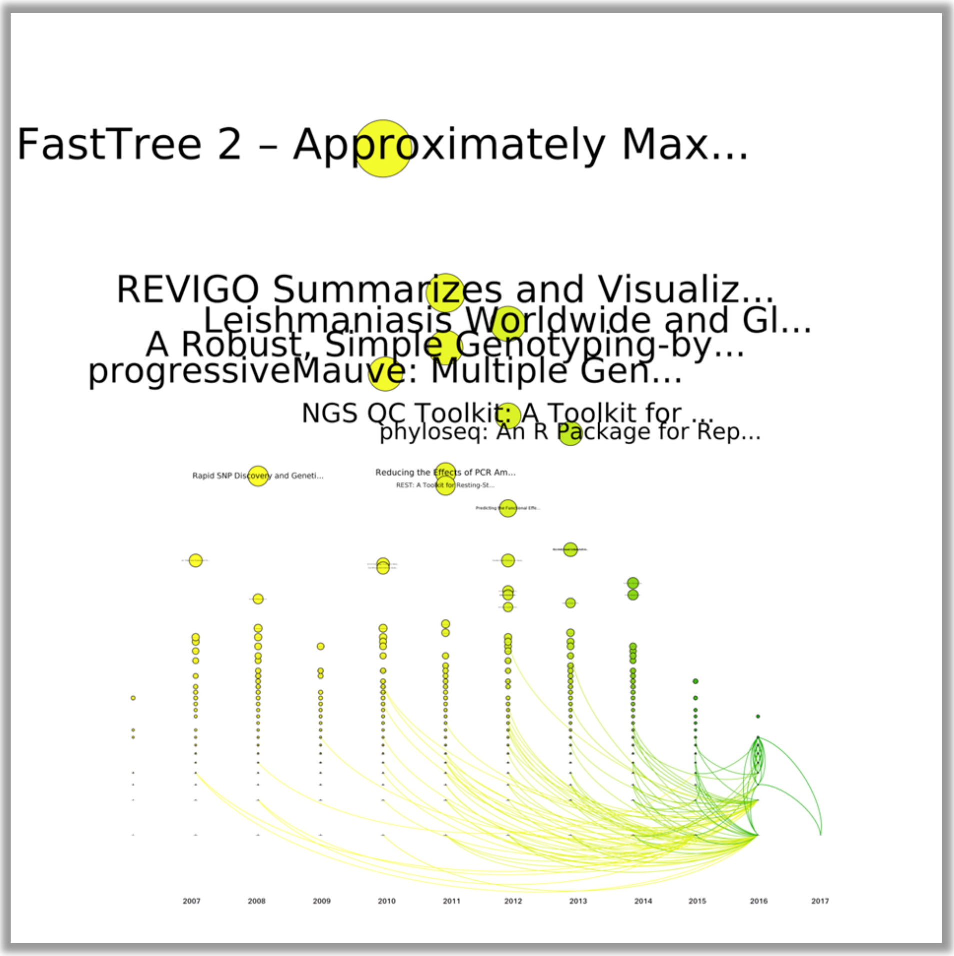


Fig. 10. The citation count ranking year-by-year. Source: own study

research on the development of science are presented in the another paper (Kamińska, 2018b).

Conclusions

The article, showing exemplary analyzes of data from the OpenCitation Corpus, shows the possibility of extracting a subset of data from this corpus in the form of CSV files describing citation graphs. It also presents the possibility of conducting bibliometric analyzes in a tool dedicated to analysis of network structures (i.e. Gephi), which extends the potential of analyzes to the possibility of making hypotheses difficult to see in data presented in traditional tabular forms.

Publishing citation data in the form of open access opens up new opportunities for analysis of the development of science. Researchers no longer limited to using commercial databases or typing bibliographic data for further analysis (Kamińska, 2017b), gain the ability to easy obtain the reliable bibliographic data. There is no doubt that the interest of scientists involved in the evaluation of science development is increasingly directed towards various open data sources such as Crossref, PubMed or OpenCitations. This is evidenced by the recently increasing number of extensions of bibliometric analytical platforms, such as modules for importing bibliographic data from the open Crossref platform implemented in VOSviewer (Van Eck, Waltman, 2017) or Gephi (Kamińska, 2018a).

The presented examples of analyzes are directed to the didactic goal and do not give rise to concrete conclusions describing scientific importance of individual articles for the development of science. The greater the range of the OCC database, and the longer its retrospective horizon becomes, the more reliable the observed dependencies will be. On the one hand, according to common opinions, that life cycle of technical sciences publications is relatively short, the extended time of publication process in the traditional model causes considerable delays. The development of the concept of open access will undoubtedly undermine this inertia as well as the possibility of quicker observation of changes occurring in the development of particular fields of science.

Bibliography

Blazegraph – REST API. (2017). Retrieved 9 September 2017, from: https://wiki.blazegraph.com/wiki/index.php/REST_API#QUERY.

Blazegraph. (2017). Retrieved 9 September 2017, from: https://www.blazegraph.com/.

Gephi makes graphs handy – CSV format. (2017). Retrieved 9 September 2017, from: https://gephi.org/users/supported-graph-formats/csv-format/.

Kamińska, A.M. (2017a). OpenCitations – otwarty indeks cytowań publikacji naukowych. *Biuletyn EBiB*, 176. Retrieved 9 September 2017, from: http://open.ebib.pl/ojs/index.php/ebib/article/view/551.

Kamińska, A.M. (2017b). Tam, gdzie zaczyna się bibliometria, czyli jak pozyskać materiał analityczny z autopsji. *Biuletyn EBiB*, 173. Retrieved 9 September 2017, from: http://open.ebib.pl/ojs/index.php/ebib/article/view/534.

Kamińska, A.M. (2017c). Wizualizacje wybranych wskaźników bibliometrycznych na przykładzie bibliograficznej bazy danych CYTBIN. *Toruńskie Studia Bibliologiczne*, *2*(19), 163-187.

Kamińska, A.M. (2018a). ScientoMiner ICR – moduł importu danych bibliograficznych z zasobów Crossref dla platformy Gephi. *Zagadnienia Informacji Naukowej*, (1), 96-113.

Kamińska, A.M. (2018b). The application of methods of social network analysis in bibliometrics and webometrics. Measures and tools. *Nowa Biblioteka. Usługi, Technologie Informacyjne i Media*, *2*(29), 29-46.

Kamińska, A.M. (2018c). Visualizations of the GRUBA bibliographic database: From printed sources to the maps of science. In: G. Osiński, V. Osińska (eds.), *Information Visualization Techniques in the Social Sciences and Humanities* (pp. 151-174). Hershey, PA: IGI Global, Global.

Peroni, S., Dutton, A., Gray, T., Shotton, D. (2015). Setting our bibliographic references free: towards open citation data. *Journal of Documentation*, *71*(2), 253-277. Retrieved 9 September 2017, from: http://speroni.web.cs.unibo.it/publications/peroni-2015-setting-bibliographic-references.pdf.

RDF. (2017). In: *W3C*. Retrieved 9 September 2017, from: https://www.w3.org/RDF/.

SPARQL Query Language for RDF. (2017). In: *W3C*. Retrieved 9 September 2017, from: https://www.w3.org/TR/rdf-sparql-query/.

Van Eck, N.J., Waltman, L. (2017). Visualizing freely available citation data using VOSviewer. In: *CWTS*. Retrieved 9 September 2017, from: https://www.cwts.nl/blog?article=n-r2r294.

Anna Małgorzata Kamińska

PLOS ONE – a case study of quantitative and dynamic citation analysis of research papers based on the data in an open citation index (The OpenCitations Corpus)

Abstract

This paper presents a case study showing the possibilities of conducting bibliometric analysis based on the data in the open citation index OCC (The OpenCitations Corpus). For cited papers published by the PLOS ONE data were extracted from the corpus and transformed to formats enabling bibliometric analysis in dedicated tools (spreadsheets, the Gephi visualization/computing platform). Then exemplary analysis and their visualizations were conducted for papers citation network. On examples the capabilities of query language SPARQL were also shown, as using only SPARQL, made available from the OCC platform, it is also possible to perform bibliometric analyzes based on remote OCC computing resources.

Key words: OpenCitations, citation index, bibliometrics, data sources, case study, Gephi, PLOS

Anna Małgorzata Kamińska

PLOS ONE – studium przypadku ilościowej i dynamicznej analizy cytowań prac naukowych na podstawie danych z otwartego indeksu cytowań (OpenCitations Corpus)

Streszczenie

W artykule przedstawiono studium przypadku obrazujące możliwości prowadzenia analiz bibliometrycznych na podstawie danych z otwartego indeksu cytowań – OCC (The OpenCitations Corpus). Dane dotyczące artykułów cytowanych pochodzące z czasopisma PLOS ONE wyodrębniono z całości korpusu i sformatowano do postaci umożliwiającej realizację analiz w narzędziach zewnętrznych (analizy z użyciem arkusza kalkulacyjnego i aplikacji obliczeniowo-wizualizacyjnej Gephi). W artykule zaprezentowano przykładowe badania i wizualizacje grafów cytowań artykułów, a także przedstawiono możliwości języka zapytań SPARQL, umożliwiającego prowadzenie analiz wprost na platformie OpenCitations (udostępnionej jako usługa WWW lub uruchomionej we własnym środowisku obliczeniowym).

Słowa kluczowe: bibliometria, Gephi, indeks cytowań, OpenCitations, PLOS, studium przypadku, źródła danych