Open Citation Content Data

Mikhail Kogalovsky[c], Thomas Krichel[a], Victor Lyapunov[a], Oxana Medvedeva[a], Sergey Parinov[ab], Varvara Sergeeva[a]

a) Russian Presidential Academy of National Economy and Public Administration (RANEPA), Moscow, Russia
b) Central Economics and Mathematics Institute of RAS (CEMI RAS), Moscow, Russia
c) Market Economy Institute of RAS (MEI RAS), Moscow, Russia

sparinov@gmail.com

Abstract. This paper presents results from the CyrCitEc project. One of this project aims is to create a publicly available source of open citation content data extracted from papers available at a research information system. Current results include four main outputs: 1) an open source software to parse papers' metadata and full text PDFs; 2) an open service to process papers' PDFs to extract citation data; 3) a dataset of citation data, including citation contexts (currently mostly for papers in Cyrillic); and 4) a visualization tool providing for users insight into the citation data extraction process and a some control over results of the citation data parsing.

Keywords: Open Data, Citation Content, CyrCitEc, RePEc, Socionet

1 Introduction

Currently there is a clear trend in the research community to make the citation data extracted from research papers more re-usable. One example is the OpenCitations project. Its main aim is "the creation and current expansion of the Open Citations Corpus (OCC), an open repository of scholarly citation data made available under a Creative Commons public domain dedication, which provides in RDF accurate citation information (bibliographic references) harvested from the scholarly literature" (http://opencitations.net/). As of June 1, 2018 this project provides references from 302,758 citing bibliographic resources. It contains information about 12,830,347 citation links to 6,549,665 cited resources.

The primary focus of the OpenCitation project is the references from research papers. Another part of the citation data that also available in research papers is the citation content or context.

In recent years, methods for analyzing the content of citations have been actively developed. Some studies (Ding et al., 2014) present a concept of the content-based citation analysis (CCA), which addresses a citation's value. It became a common view that "the text of citation context is used to characterize publications for various applications, such as publication summarization, survey article generation and information retrieval" (He and Chen 2017). Other authors wrote: "the extraction of citation contexts is a preliminary step to any statistical, distributional, syntactic or semantic analy-

sis" (Bertin and Atanassova 2018). Also "to capture document usage, we observe that the context in which one document cites another tends to reflect how a document is used, namely, within a document, people tend to cite other documents for very precise reasons" (Berger at el. 2017).

One of few already existing sources of open citation content data is the In-text Reference Corpus (InTeReC) available at https://zenodo.org/record/1203737. Currently the InTeReC dataset provides 314023 sentences containing in-text references (also called as the in-text citations) together with other useful data. The sentences are extracted from 90,071 research articles published by PLOS5 up to September 2013 (Bertin and Atanassova 2018).

A full text of each sentence in InTeReC (Bertin and Atanassova 2018) is supplemented by:

- a journal title;
- DOI of the article from which the sentence was extracted;
- size of the article, as number of sentences, and a position of the sentence in the article, as number of sentences from the beginning of the article;
- size of the section, as number of sentences, and a position of the sentence in the section, as number of sentences from the beginning of the section;
- section type (introduction, method, results, etc.);
- a list of verb phrases that occur in the sentence.

Our project CyrCitEc[1] provides a new source of open citation content data. The project is funded by the Russian Presidential Academy of National Economy and Public Administration (RANEPA, http://www.ranepa.ru/eng/). The first result of this project was presented in (Barrueco et al. 2017).

The project has two main aims: 1) to create the CyrCitEc system - a public service for processing available research papers full text (particularly, in PDF and with a focus on Social Sciences), in order to build and regularly update an open dataset of citation relationships and citations content; 2) to use the citation content data for developing methods of qualitative citation analysis, which can be used for improving of current practice of a research performance assessment.

The project builds a pilot version of open scholarly infrastructure (Bilder et al. 2015) based on the following pillars:

1. Open distributed architecture. The project provides a concept, open source software[2] and an initial core infrastructure for interoperable systems, which are processing citation relationships and its content from research papers' full text.
2. Two initial nodes of this core infrastructure, presented by interacting CitEc (http://citec.repec.org/) and CyrCitEc systems. Currently these nodes are exchanging by citations data. The nodes have a specialization on processing papers in specific languages: Romano-Germanic languages by CitEc and Russian by CyrCitEc. Other nodes, e.g. specialized on processing citation data in languages, like Chinese, Japanese, Arabic, etc., could be added by the same way. There is also an intention to integrate data about references into the OpenCitations Corpus (http://opencitations.net/).

[1] https://github.com/citeccyr/CyrCitEc_method

[2] https://github.com/citeccyr

3. Transparency. It allows publishers, authors and readers of papers to see how the citation data of their papers are extracted by the system and to trace why some papers' references / in-text citations are not processed or not counted.
4. Better representation and usability of citation data by its deeper integration with research information system tools and services.
5. Enrichment facilities. The system provides tools for authors of papers to enter additional data to correct errors of processing citations found in their papers and to enrich their citation relationships, e.g. by qualitative characteristics of their motivation for citing papers of other authors, etc.
6. Public control. Readers of papers can see how authors used enrichment facilities to increase their number of citations. Thus, they will be able to react to authors misbehaviour.

CyrCitEc takes papers' metadata from the Socionet (https://socionet.ru/), which also includes a full set of metadata from RePEc (http://repec.org).

Comparing with InTeReC, the CyrCitEc system has following main differences:

a) an openness for adding new papers for processing by the system. The papers just have to be added to a Socionet or into RePEc;
b) the system works as a part of an infrastructure, i.e. in everyday mode it automatically processes all new available papers and updates citation content data;
c) the input papers are in PDF (InTeReC works with papers in XML).

Unlike of InTeReC authors using the term "in-text reference", we in CyrCitEc use the term "in-text citation". It is defined as: "the in-text citations of publications are the citations referred to this publication in the full text of other publications cited this publication. The text around the in-text citation is the citation context text" (He and Chen 2017).

The second section of the paper presents the open citation content data provided by the CyrCitEc project for public re-use.

In the third section, we describe a visualization tool of the citation content data which creates a transparence on how the citation data is produced step by step. It also allows some public control over results of citation data parsing.

The last section briefly discusses possible further development of this project.

2 Open Citation Content Data from the CyrCitEc project

In the beginning of June 2018, CyrCitEc processed 220 collections of papers with 100,553 publications in total (see Tab. 1 for more statistics). The biggest part of this set are 157 Russian academic journals covering different academic disciplines and provided by the NEICON consortium. There are also research papers series in Russian and English languages provided by Russian Universities (RANEPA, Higher Scholl of Economics, etc.) and by research organizations of Russian Academy of Sciences.

An approach used by CyrCitEc for the citation content data parsing was presented in (Parinov 2017).

All extracted by CyrCitEc project citation data and processing log files are publicly available at http://peren.openlib.org/. This storage is organized as nested folders with names based on Socionet IDs of processed papers. A processed paper's folder contains: a) JSON version of PDF papers (the file 0.pdf-stream.json), which was used for parsing citation data; b) file "summary.xml" with the parsed citation data; and c) reports about errors in processing the paper and parsing citation data (files with extensions ".err" and ".log").

A single in-text citation includes following data:

1) a text string of how this in-text citation is occurred in a paper content, e.g. a number or an author name in square or round brackets (the tag `<Exact>` in the example below);

2) a link to a reference, mentioned in this in-text citation (the tag `<Reference>` below);

3) text coordinates of the in-text citation, i.e. a serial number of the first and the last in-text citation symbols counting from the begging of the paper's content (tags `<Start>` and `<End>`);

4) citation contexts located at the left and at the right according the in-text citation; it includes at least 200 symbols expanded for taking a whole sentence (tags `<Prefix>` and `<Suffix>`).

An example of parsed data for one in-text citation:

```
<intextref>
  <Prefix>… countries and Soviet republics</Prefix>
  <Suffix>; Gokhberg, Kuznetsova, 2011]. …</Suffix>
  <Start>8757</Start>
  <End>8781</End>
  <Exact>[Gokhberg et al., 2009</Exact>
  <Reference>20</Reference>
</intextref>
```

Source: https://goo.gl/1FAkCH

The in-text citation from the example above has a link with a reference having the number 20 in the paper. CyrCitEc parsed for this reference following data:

```
<reference num="20" start="54464" end="54654"
  author="Gokhberg Kuznetsova …" title="Towards …"
  year="2009"
  handle="repec:oup:scippl:v:36:y:2009:i:2:p:121-126">
  <from_pdf>Gokhberg L., Kuznetsova T., Zaichenko
    S. (2009) Towards a New Role of Universities in Rus-
sia:
    Prospects and Limitations. Science and Public Policy,
    vol. 36, no 2, pp. 121-126.</from_pdf>
</reference>
```

Source: https://goo.gl/1FAkCH

XML data of the example above includes following subtags and attributes:

a) subtag `<from_pdf>` - extracted raw data of a reference (some publishers provide reference data within the papers' metadata, see the subtags `<from_metadata>` in the example at https://goo.gl/tKRZF1);

b) attribute `num` - a serial number of the reference in the paper's list of references;

c) attributes `start` and `end` - text coordinates of the reference, which are numbers of the first and the last symbols of the reference counted from the beginning of the initial PDF document's text;

d) attribute `url` – contains a proper URL, if there is one in data of the tag `<from_pdf>`;

e) attributes `author`, `title` and `year` are extracted from the raw reference data in the tag `<from_pdf>` and used for different purposes, e.g. for searching the in-text citations at Socionet for linking the reference with metadata of the same paper (creating a citation relationship for this reference), etc.;

f) attribute `handle` – contains ID of the paper at Socionet, if the linking procedure for this reference was successful.

These data about in-text citations and references are supplemented by the ID of paper's metadata (see `<source handle=` in the example below) and by the URL of the source full text PDF of the paper (see `<futli url=` below). Having the paper's metadata ID and using Socionet API one can take all available information about this paper, including its title, abstract, authors, etc.

```
<source handle="repec:hig:fsight:v:11:y:2017:i:4:…">
<futli url="https://foresight-journal.hse.ru/data/…">
```

Source: https://goo.gl/1FAkCH

Comparing with InTeReC, the CyrCitEc data have following main differences:

1. the citation content is organized as two text strings: at the right and at the left side according the in-text citation location; and it can provides several sentences instead of one sentence in InTeReC;
2. a broader set of attributes for citation content, like reference data linked with in-text citation, etc.
3. in-text citation's coordinates as number of symbols (InTeReC counts sentences);
4. current version of CyrCitEc citation data has no associations with the type of paper's sections that exists in InTeReC.

Citation data generated by CyrCitEc system provides more opportunities for a citation content analysis and allow different types of visualization of this data, one of which is presented below.

3 Visualization tool for the open citation content data

For open citation content data there is a visualization tool, which presents daily updated aggregated statistics about parsing results for each collection of papers. The tool is publicly available at http://citru.repec.org/stats.html.

series	[1] records	[2] with futli	[3] with WARCs	[4] with PDF WARCs	[5] with JSON
RePEc:bkr:wpaper	26	26	21	17	17
RePEc:cas:wpaper	26	26	20		
RePEc:cfr:cefirw	228	228	170	164	160
RePEc:eer:wpalle	240	240	194		
RePEc:eus:ce3swp	26	26	23	23	23
RePEc:eus:wpaper	53	53	42	42	40
RePEc:gai:gbchap	37	37	37	36	36

Fig. 1. Visualization tool's main page, a fragment (source: http://citru.repec.org/stats.html)

The main page of the visualization tool (see Fig. 1) provides a general overview of processed data by collections of papers (collections in rows). For each collection it presents three groups of statistics: a) numbers of papers processed on different stages of the CyrCitEc utilization (columns [1]-[5]); b) numbers of papers' JSON versions which contain (or not) some citation data (columns [6]-[9]); c) numbers of found (or not) different types of citation data (columns [a]-[e]).

Table 1 presents some data from this page (on 2018.06.01): 1) the column "Totals" - aggregate statistics about different aspects of processing of research papers from all collections; 2) the column "Collection" - statistics for all papers of a sample collection (citation data used in the previous section belong to a paper from this collection).

Table 1. Aggregate statistics of the open citation content data (on 01.06.2018)

Column	Legend	Totals	Collection
series	list of processed collections of papers	220	RePEc:hig:fsight
[1]	metadata records available	100553	262
[2]	records with links to paper's full text	91929	262
[3]	records at Web ARChive (WARC)	86882	216
[4]	PDF files in Web ARChive	67048	213
[5]	PDF files converted to JSON	66802	207
[6]	JSON files with found reference sections	49129	95
[7]	JSON with in-text citations	45580	94
[8]	JSON files with citation relationships	18977	75
[9]	JSON files with non-mentioned references	18617	92
[a]	total references	865042	3455
[b]	total citation contexts	735709	4041

[c]	total mentioned references	806680	2700
[d]	total citation relationships	47052	597
[e]	total non-mentioned references	129333	755

Tab. 1 provides also legends for the main page's all columns. The legends for columns [1]-[5] explain functions of the CyrCitEc utilization stages. The first two stages are: taking all available papers' metadata from Socionet and specify which of them have a link to a paper's full text. At the stage "[3] records at Web ARChive" the system uploads available papers' full texts into a web archive based on the WARC archive format[3]. The next stage "[4] PDF files in Web ARChive" selects only PDF from all available papers' full text formats. The stage "[5] PDF files converted to JSON" determines which PDFs have a text layer and are not corrupted. After conversion only these files have a proper JSON format and can be processed further for the citation data parsing.

The numbers in columns [6]-[9] of the main page show following amounts of:

- papers with recognized list of references (49,129 of 66,802 are available for analysis);
- papers with recognized in-text citations (45,580 of 49,129 papers have recognized in-text citations);
- papers with references which we linked with metadata of the same papers available at Socionet (18,977 of 49,129 papers have at least one citation relationship);
- papers with non-mentioned references (18,617 of 49,129 papers have non-mentioned references).

The main page's columns [a]-[e] contain total numbers of parsed records:

- references ([a]);
- citation contexts, which include at least one in-text reference ([b]);
- references for which there are at least one in-text reference ([c]);
- references linked with metadata of the same paper ([d]); and
- references without mentions in text ([e]).

The column "series" (left at Fig. 1) contains ID of collections, which currently have being processed by CyrCitEc. If some ID in the column has no link, it means that in this collection there are no papers with recognized list of references. In such cases the system cannot produce the citation statistics for this collection. If the link exists, it opens a page with a list of citation data by the collection's papers. Tab. 2 shows an example of statistics for a single paper in the list. A paper for this example is the same which citation data is used in the previous section.

Table 2. Aggregate citation statistics for a single paper (on 01.06.2018)

paper handle	RePEc:hig:fsight:v:11:y:2017:i:4:p:84-95
[a] references	40
[b] reference contexts	45
[c] contexts by reference	34

[3] https://en.wikipedia.org/wiki/Web_ARChive

[d] linked references	7
[e] references without contexts	6

For each paper, presented by its ID (see the "paper handle" in Tab. 1), the system provides following data:

- number of recognized references with a link to a list of them;
- number of recognized citation contexts, which should include mentions of one or more references;
- number of references mentioned in a paper;
- number of references with available metadata of the same papers;
- number of non-mentioned references in a paper.

All values on this page have links to further details. E.g. a link from paper's ID opens paper's metadata page at Socionet. Links from numbers at this page open list of appropriate entities, like recognized references, extracted citation contexts, etc.

Fig. 2 presents a fragment of a list of non-mentioned references of a paper. Each non-mentioned reference in the list have a hyperlink called "check this in PDF content". This link opens the paper's PDF where the non-mentioned reference is highlighted as an annotation. One can check is the reference really not mentioned and if it is a technical error he/she can report about this to initiate improvements of citation data parsing software.

The 6 references without contexts in paper RePEc:hig:fsight:v:11:y:2017:i:4:p:84-95

17

Gertner J. (2012) The Idea Factory: Bell Labs and the Great Age of American Innovation, New York: Penguin Group.
(check this in PDF content)

24

Kuzyk M., Grebenyuk A., Kakaeva E., Manchenko E., Dovgiy V., pp. 84–95
(check this in PDF content)

26

Fig. 2. A fragment of non-mentioned references list

Columns [1]-[5] of the main page (see Fig. 1) also provides links to papers missed at different stages of the CyrCitEc processing. Tab. 4 shows an example of how numbers of papers are changed step by step of their processing by the system.

Table 4. Numbers of processed papers for a collection (on 01.06.2018)

Series	[1]	[2]	[3]	[4]	[5]
RePEc:hig:fsight	262	262	216	213	207

The same numbers in columns [1] and [2] means that all available papers metadata have links to papers' full text. A number in the column [3] is less than the number in [2] since for 42 papers the system cannot download them for parsing citation data. The column [4] has number in 3 papers less than in [3] what means that these 3 papers are not PDF. A difference between numbers in the column [5] and [4] means that 6 papers cannot be converted to JSON format for further processing, since, e.g. the papers have no text layer or corrupted.

For all cases when there are differences between numbers in neighbor columns the system provides links to pages with list of missed papers. Using this feature one can see which papers were out of processing by CyrCitEc and figure out why.

4 Conclusion

The CyrCitEc project provides for public re-use four main outputs: 1) the open source software to parse and manage citation data; 2) the open service to process paper's PDFs to extract citation data; 3) the dataset of citation data, including citation contexts; and 4) the visualization tool to provide for users a transparence on how extraction of citation data works and a public control over results of parsing citation data.

References

1. Barrueco, J. M., Krichel, T., Parinov, S., Lyapunov, V., Medvedeva, O., & Sergeeva, V. (2017). Towards Open Data for the Citation Content Analysis. *arXiv preprint arXiv:1710.00302*.
2. Berger, M., McDonough, K., & Seversky, L. M. (2017). cite2vec: citation-driven document exploration via word embeddings. IEEE transactions on visualization and computer graphics, 23(1), 691-700.
3. Bertin, M., & Atanassova, I. (2014). A study of lexical distribution in citation contexts through the IMRaD standard. PloS Negl. Trop. Dis, 1(200,920), 83-402.
4. Bertin, M., & Atanassova, I. (2015). Factorial Correspondence Analysis Applied to Citation Contexts. In BIR@ ECIR (pp. 22-29).
5. Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. Journal of the Association for Information Science and Technology, 67(1), 164-177.
6. Bertin, M., & Atanassova, I. (2018). InTeReC: In-text Reference Corpus for Applying Natural Language Processing to Bibliometrics. In Proc. of the Seventh Workshop on Bibliometric-enhanced Information Retrieval (BIR), Grenoble, France, CEURWS.org (pp. 54-62).
7. Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. Journal of Informetrics, 12(1), 59-73.
8. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content□based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology, 65(9), 1820-1833.

9. He, J., & Chen, C. (2017). Understanding the changing roles of scientific publications via citation embeddings. arXiv preprint arXiv:1711.05822.
10. Hernández-Alvarez, M., & Gómez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering, 22(3), 327-349.
11. Parinov, S. (2013). Towards a Semantic Segment of a Research e-Infrastructure: necessary information objects, tools and services. International Journal of Metadata, Semantics and Ontologies 6, 8(4), 322-331. https://socionet.ru/publication.xml?h=repec:rus:mqijxk:32
12. Parinov, S. (2017). Semantic Attributes for Citation Relationships: Creation and Visualization. In Research Conference on Metadata and Semantics Research (pp. 286-299). Springer, Cham.
13. Parinov, S., Lyapunov, V., Puzyrev, R., & Kogalovsky, M. (2015). Semantically enrichable research information system SocioNet. In International Conference on Knowledge Engineering and the Semantic Web (pp. 147-157). Springer, Cham.