Open Citation Content Data and a Digital Library

© Sergey Parinov
Central Economics and Mathematics Institute of RAS,
Moscow, Russia
Russian Presidential Academy of National Economy and Public Administration,
Moscow, Russia
sparinov@gmail.com

Abstract. Typically, a digital library serves as a source of research papers for extracting of citation content data from papers' full text. When a digital library can also be a recipient of these data, it opens new ways to make the knowledge more open. This paper presents some results of an integration of the open citation content data, provided by CyrCitEc project, into the big research digital library Socionet. Additionally to citation relationships, papers in Socionet also have linkages with authors' personal profiles and through them with other information objects. It allows making an enrichment of data for the citation content analysis by different additional information and, as well, linking results of such analysis with objects in a digital library, like papers, their authors, affiliation organizations, etc. A digital library can visualize results of the citation content analysis as computer-generated annotations to full texts in PDF. These results provide further development of a research e-infrastructure particularly by sharing, managing and curating citation content data, making them FAIR.

Keywords: Citation Content, CyrCitEc, Digital Library, Socionet, Word2Vec

1 Introduction

Currently there is a clear trend in the research community to make more re-usable citation data from research papers. One of illustrations is the OpenCitations project. The main aim of this project 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 April 10, 2018 this project makes open the references from 302758 citing bibliographic resources, it contains information about 12830347 citation links to 6549665 cited resources.

The focus of the OpenCitation project is the references. Another part of citation data that also available in research papers is the citation content or context. Waltman (2016) in his review of the traditional citation impact indicators proposed different ways for the indicators' improvement, including taking into account "the context in which a publication is referenced (i.e., the sentences in a citing publication around the reference to a cited publication)" (Waltman 2016 p. 43).

In recent years, methods for analyzing the content of citations have been actively developed. Some studies (Zhang et al., 2013; Ding et al., 2014) present a concept of the content-based citation analysis (CCA), which addresses a citation's value. "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 analysis" (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).

Practical experiments with the extraction and analysis of the in-text citations (they are also called as the in-text references) on various sets of full text papers are also known. One of them identified verbs in citation contexts (Bertin and Atanassova, 2014). "Our hypothesis is that the semantic meaning of the relation that exists between the cited work and the citing article is often expressed, to some extent, by the verb phrase in the sentence containing the in-text reference" (Bertin and Atanassova 2018). These authors also characterized the different sections of articles in terms of the verbs that appear in citation contexts (Bertin and Atanassova, 2015).

Another aspect of the citation content analysis is how references are distributed along the structure of articles, as well as the age of these cited references (Bertin et al., 2016). Some authors analyzed in-text citations as functions of time, textual progression, and scientific field. They built characteristics of the in-text citations in over five million full text articles (Boyack et al., 2018).

Hernández-Alvarez and Gómez (2016) in their survey of citation context analysis provided information about used tasks, techniques, and resources, including such tasks as the citation polarity and function classifications.

The analysis of citation polarity/function has a potential for conclusions with some accuracy about the

Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), Moscow, Russia, October 9-12, 2018

motives of authors to cite papers. Such analysis can also produce suggestions: what exactly from the cited papers and for what purposes were used in the citing papers. In some cases, this information may be critically important to authors of the cited papers and may help to initiate direct communication between them and citing authors.

The second section of the paper presents the CyrCitEc project and its current results, which aimed to provide an open citation content data source.

In the third section, we describe some services of the digital library Socionet that use the citation content data. One of them gives better representation of citations in papers' full text PDF. Another, a network of semantic linkages among information objects in Socionet and its API create an opportunity to enrich citation content data by additional information.

The last section discusses a preparation of the citation content datasets for building Word2Vec models. These models allow analysis of similarities between words from the citation content and IDs of citing/cited papers and IDs of personal profiles of citing/cited authors. The section provides links to created models.

2 Open citation content data

One of few already existed 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 is supplemented by (Bertin and Atanassova 2018):

- 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.

Another source of open citation content data is provided by our ongoing project CyrCitEc (https://github.com/citeccyr). The project head is Oxana Medvedeva. It is funded by the Russian Presidential Academy of National Economy and Public Administration (RANEPA, http://www.ranepa.ru/eng/).

The project has two main aims: 1) to create a public service for processing available research papers full text (particularly, in PDF and with main 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 a current practice of a research performance assessment.

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

1. Open distributed architecture. It means providing a concept, open source soft-ware and an initial core infrastructure for interoperable systems, which are pro-cessing citation relationships and 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., can be added by the same way. There is also an intention to integrate data about references into the OpenCitations Corpus (http://opencitations.net/).
3. Transparency. It allows publishers, authors and readers of papers to see for each paper how their citation data are created 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 a digital library tools and services.
5. Enrichment facilities. The system should provide tools for authors of papers to enter additional data to correct errors while processing citations from 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 should see how authors used enrichment facilities to increase their number of citations. Public will be able to react on wrong authors behaviour.

CyrCitEc takes papers' metadata from the Socionet digital library (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:

- an openness for adding new papers for processing by the system, the papers just have to be added to a digital library Socionet or into RePEc;
- the system in everyday mode automatically processes all new papers at its in-put and daily updates citation content data;
- the input papers are in PDF (InTeReC works with papers in XML).

In CyrCitEc we use the term "in-text citation" instead of the "in-text reference" accepted in InTeReC. "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).

In the beginning of April 2018, CyrCitEc processed 203 collections of papers with 89342 publications in total. The biggest part of this set are 157 Russian

academic journals covering different academic disciplines and provided by NEICON (https://socionet.ru/collection.xml?h=spz:neicon&l=en). There are also research papers series in Russian and English languages provided by the Higher Scholl of Economics (https://socionet.ru/collection.xml?h=repec:hig&l=en), RANEPA (https://socionet.ru/collection.xml?h=repec:rnp&l=en) and some other Russian Universities.

An approach used by CyrCitEc for citation data parsing was presented in (Parinov 2017). Victor Lyapunov and Sergey Petrov built all needed software to parse citation data from PDF documents.

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. E.g. the folder http://peren.openlib.org/RePEc/hig/fsight/v%253A11%253Ay%253A2017%253Ai%253A4%253Ap%253A84-95/ contains:

a) JSON version of PDF papers (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").

Daily updated aggregated statistics about parsing results for each collection are available at http://cirtec.ranepa.ru/stats.html. Thomas Krichel maintains these statistics.

Processing statistics of this set of collections show (on the end of May 2018) that only 69% of total papers have full text PDF available for the citation data parsing and only 51% of total papers have a list of references in more or less standard form.

Based on the subset of papers with references we parsed in total 801318 references that is in average 18 references per paper. In this set, we have about 5% of duplicated references, because different papers cite the same publications and have the same references.

For 26467 of parsed references we were able to create citation relationships between citing and cited papers, since we found cited papers' metadata within Socionet.

Additionally, we parsed 750607 in-text citations. They mention 1072175 of parsed references. It is in 270857 references more than total number of parsed references, since some references are mentioned more than one time. In average, it is 1.3 mentions per reference.

Non-mentioned references were also counted: 110340 references (it is 14% of total) have no mentions in the in-text citations at all. About 37% of papers with references have at least one non-mentioned reference.

One in-text citation includes following data (see also an example of the data below):

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 about 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 a 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:20
09:i:2:p:121-126">
  <from_pdf>Gokhberg L., Kuznetsova
T., Zaichenko S. (2009) Towards a New
Role of Universities in Russia:
Prospects and Limitations. Science
and Public Policy, vol. 36, no 2, pp.
121-126.
  </from_pdf>
</reference>
```

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

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

a) subtag <from_pdf> - extracted raw data of a reference (some publishers provided reference data within papers' metadata, see as an example any citation data file of NEICON archive);

b) attribute num - a serial number of the reference in the list;

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 row reference data in the tag `<from_pdf>` and used for different purposes, e.g. for searching in-text citations by author names, 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 digital library, 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 in a digital library (see `<source handle=` in the example below) and by the URL of the source full text PDF of the paper (see `<futli url=` below). Using the paper's metadata ID one can have 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 source has following main differences:

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

3 Digital library and citation content data

The citation data provided by the CyrCitEc project include ID of source papers and URL of their full texts (see the last example above). Such links allow us, using API of a digital library like Socionet (Parinov et al. 2015), to provide new features for users of the digital library and to enrich citation content analysis, as well. In general, it opens new ways to make the knowledge more open.

Socionet services, as it is described in (Parinov 2017), use in-text citations and references data to produce computer-generated annotations to the content of PDF papers. Fig. 1 shows how these annotations look like using the in-text citation and the reference from the examples above.

Readers of PDF papers see the in-text citations, if they exist, as an annotated text. At Fig. 1 there are mentions of two references in brackets. These highlighted in-text citations works as interactive elements, since a click on them opens an information box (at the right side on Fig. 1) with additional data about the cited paper. The additional data can include details about the cited paper (citing statistics, title, authors, etc.) and links to some tools.

Another Socionet feature is the multiple semantic relationships between information objects (Parinov 2013). It allows an enrichment of citation content analysis by expanding the citation data with additional attributes belonged to semantically linked information objects. A fragment of the semantic linkage network existed at Socionet is presented at Fig. 2.

Using these linkages, we can associate with the citation data different additional data of cited and citing papers, like titles, authors, classification codes and other elements of their metadata.

Currently, Socionet already has about 70000 authors' profiles. These authors' profiles have linked with authors' papers, and with about 15000 profiles of organizations. The organizations' profiles also have links with other authors' profiles belonged to their staff.

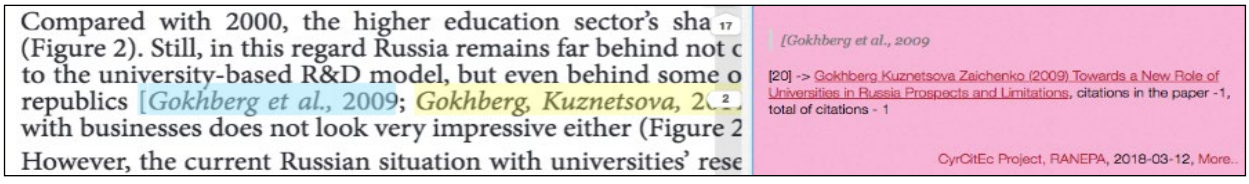
Socionet API allows taking different data related with selected information object specified by its ID:

1. Paper's metadata including list of semantically linked objects. The in-text citation and reference examples from the previous section have, as a source, a paper with the ID: repec:hig:fsight:v:11:y:2017:i:4:p:84-95. To take the paper's metadata one should run API: https://socionet.ru/fs/ap.cgi?h=repec:hig:fsight:v:11:y:2017:i:4:p:84-95. The output is a metadata in XML, which additionally to usual bibliographic data about a paper data includes authors' profile ID (see within the API output the tag `<coauthor handle="repec:per:pers:pku327"…`), paper's references (`<out><citation>`), paper's in-text citations in a form of annotations to full text PDF (`<annotations><annot>`), and many other data.

2. Author's personal profile data including linked objects, like author's papers, affiliation organization, etc. To take author's profile data, which ID is specified in the example above, run the API: https://socionet.ru/fs/ap.cgi?h=repec:per:pers:pku327. It output includes a list of papers' ID, which the author claimed as own, and the author's workplace ID (`<workplace><link handle="repec:edi:aneeeru"…`).

3. Organization's profile data including links with authors who work for it. API works here similarly as in previous case.

Fig. 1. An in-text citation as an interactive element (source: https://goo.gl/bZJwzZ)



By this way, one can aggregate various sets of enriched citation data for different scenarios of the citation content analysis. One can collect the citation contents from all papers for a specified author to analyze how the author cite other papers. Similar collection of data can be created to analyze how the author is cited by other researchers. Results of such analysis can be aggregated for groups of authors, e.g. who works for the same organization, etc.

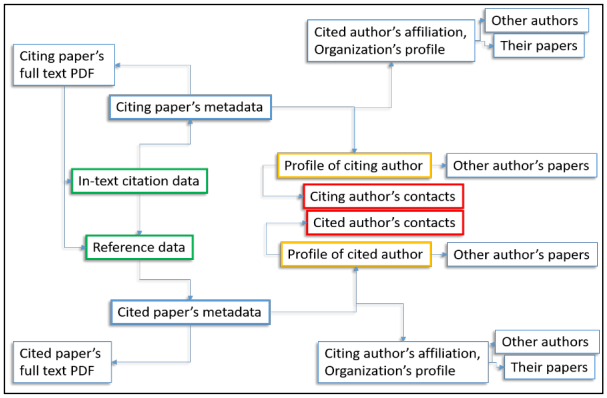


Fig. 2. Semantic linkage network in a digital library, a fragment

4 Discussion

Using the method to enrich initial citation content by data about citing and cited papers/authors, we get new opportunities for the citation content analysis.

We have the citation content as a text provided by pairs of tags <Prefix> and <Suffix> (see the first example in the section "Open citation content data"). The idea is to add into this text IDs of the citing/cited papers and IDs of its authors' profiles.

As a result we have a dataset of text strings with embedded data about all main entities related to citations. It allows an analysis of words (or IDs) that occur in similar contexts. One of possible methods for such analysis is the word embedding approach, which is able to capture multiple different degrees of similarity between words (Mikolov et al. 2013).

This approach, implemented as Word2Vec method and software (https://code.google.com/archive/p/word2vec/), has become popular for some types of the citation content analysis. He and Chen (2017) used this method "to provide temporal representations of in-text citation of publications by word embedding models trained from citation context text, which can be used to characterize and analyze the changing and complex roles of the publications" (He and Chen 2017). Another scholars used Word2Vec to create the cite2vec method: "Our approach projects words that are representative of documents into a 2D space, and subsequently projects documents into this same space such that a document's proximity to a word indicates its manner of usage, while also preserving similarity of document-to-document usage" (Berger at el. 2017). They wrote: "cite2vec visualizes documents and words in a way that adheres to such citation contexts, so that the user can explore and discover documents in a usage-oriented manner" (Berger at el. 2017).

Our aim is to analyze similarities between words of the citation content and IDs of citing/cited papers and authors. By this way we can specify typical contexts that authors use for citing papers and also the contexts which papers and authors are cited in.

To create the input dataset for Word2Vec we used all citation contents from the tags <Prefix> and <Suffix>, provided by CyrCitEc project, excluding papers in English (at the current stage we analyze Russian papers only). The dataset contains 759922 text strings.

We enriched this dataset by adding ID of the citing paper into each citation content text string. We split up the ID on four part. It allows us to analyze similarities for following cases: a) the selected citing paper; b) all citing papers of the selected journal; c) all citing papers of the selected publisher.

This dataset is available at http://cirtec.ranepa.ru/Word2Vec/fixes.raw.txt.

For this dataset we created the Work2Vec models (both CBOW and Skip-gram types) with following parameters:

```
-size 200 -window 5 -sample 1e-4 -
negative 5 -hs 0 -binary 1 -cbow 1 -
iter 3.
```

The models are available at http://cirtec.ranepa.ru//Word2Vec/ in files: fixes.raw.cbow.bin and fixes.raw.skig.bin. They had following characteristics (on the end of May 2018): Vocab size: 665817; Words in train file: 64207538; Alpha: 0.000005; Progress: 100.00%. We use this model as a base for comparisons.

The next version of the input dataset for Word2Vec has the citation contents processed by morphological analyzer for Russian language. We use the Pymorph2 package (https://github.com/kmike/ pymorphy2) to replace words of the initial citation content text by its normal forms. By this we removed variations of words. This second dataset is available at http://cirtec.ranepa.ru/Word2Vec/fixes.stem.txt.

The same two types of models are available at http://cirtec.ranepa.ru/Word2Vec/ if files: fixes.stem.cbow.bin and fixes.stem.skig.bin. They had following characteristics (on the end of May 2018): Vocab size: 191137; Words in train file: 55954677; Alpha: 0.000005; Progress: 100.00%.

Comparing characteristics of these two sets of models one can see that a removing of variations of words decreased the number of words in the vocabulary about 3 times less and in the train file – about 8 millions.

Using the method illustrated by the Fig. 1 we can integrate into the Socionet results of the Word2Vec based citation content analysis. Since we can identify for each word in the Word2Vec models from what citation content of what paper it comes, we can visualize information about found similarities for this word as a computer-generated annotation to the word in the paper's full text PDF. Readers of the paper will have more

information related to the paper's content and this makes the knowledge more open.

All data described above are freely available for re-use and analysis. Currently, it is for papers in Russian language. We plan further enrichment of the initial citation contents and to make more versions of the Word2Vec models with focus on different types of similarities based on papers in Russian and, as well, in English (RePEc papers). Results of using the Word2Vec models for analysis of similarities will be published later.

Acknowledgments. A part of this research – the approach development for extracting citation content data with focus on the supercomputer simulation of interactions among the agents and research community environment is funded by RSF grant (project No. 14-18-01968).

References

[1] 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.

[2] 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.

[3] Bertin, M., & Atanassova, I. (2015). Factorial Correspondence Analysis Applied to Citation Contexts. In BIR@ ECIR (pp. 22-29).

[4] 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.

[5] 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).

[6] 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.

[7] 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.

[8] He, J., & Chen, C. (2017). Understanding the changing roles of scientific publications via citation embeddings. arXiv preprint arXiv:1711.05822.

[9] Hernández-Alvarez, M., & Gómez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering, 22(3), 327-349.

[10] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

[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.