Analysing Structured Scholarly Data Embedded in Web Pages

Pracheta Sahoo[2], Ujwal Gadiraju[1], Ran Yu[1], Sriparna Saha[2], and Stefan Dietze[1(✉)]

[1] L3S Research Center, Leibniz Universität Hannover, Hannover, Germany
{gadiraju,yu,dietze}@L3S.de
[2] Indian Institute of Technology, Patna, India
{pracheta.mtmc14,sriparna}@iitp.ac.in

Abstract. Web pages increasingly embed structured data in the form of microdata, microformats and RDFa. Through efforts such as schema.org, such embedded markup have become prevalent, with current studies estimating an adoption by about 26% of all web pages. Similar to the early adoption of Linked Data principles by publishers, libraries and other providers of bibliographic data, such organisations have been among the early adopters, providing an unprecedented source of structured data about scholarly works. Such data, however, is fundamentally different from traditional Linked Data, by being very sparsely linked and consisting of a large amount of coreferences and redundant statements. So far, the scale and nature of embedded scholarly data on the Web has not been investigated. In this work, we provide a study on embedded scholarly data to answer research questions about the depth, syntactic and semantic characteristics and distribution of extracted data, thereby investigating challenges and opportunities for using embedded data as a structured knowledge graph of scholarly information.

Keywords: Linked Data · Scholarly articles · Web Data Commons · Analysis

1 Introduction

Bibliographic data is widespread on the Web. Libraries and publishers have in particular embraced the Linked Data principles and associated W3C standards throughout the past decade, making large amounts of bibliographic metadata available on the Web [2]. However, uptake and reuse is still hindered by a variety of issues, including the lack of dynamics, and to a certain degree, scale.

More recently, annotations embedded in HTML pages have become another prevalent source of structured data on the Web, building on standards such as RDFa[1], Microdata[2] and Microformats[3]. Markup is used by search engine

[1] RDFa W3C recommendation: http://www.w3.org/TR/xhtml-rdfa-primer/.
[2] http://www.w3.org/TR/microdata.
[3] http://microformats.org.

A. González-Beltrán et al. (Eds.): SAVE-SD 2016, LNCS 9792, pp. 90–100, 2016.
DOI: 10.1007/978-3-319-53637-8_10

providers to interpret content of Web pages or enrich result pages with factual entity descriptions. One central effort is the schema.org initiative[4], driven by Google, Yahoo!, Bing and Yandex, aiming at defining a common vocabulary for describing entities on the Web and driving its adoption. A recent initiative [3] investigating a large-scale Web crawl from 2014 of 2.01 billion HTML pages constituting more than 15 million pay-level-domains (PLDs) found that 26% of all pages contain some form of embedded markup already, resulting in a corpus of 20.48 billion RDF quads[5].

Considering the apparent upward trend of adoption [1] (the proportion of pages containing markup increased from 5.76% to 26% between 2010 and 2014) and the still comparably limited nature of the investigated Web crawl, the scale of the data suggests a significant potential for exploiting it for a wide range of tasks, such as entity retrieval, knowledge base population or entity summarization.

Despite a growing interest in such embedded semantics, a thorough understanding of its adoption for scholarly resource metadata is still lacking. In this paper, we present the first study of scholarly data extracted from embedded annotations, utilizing the Web Data Commons as the largest crawl of embedded markup so far. Our analysis investigates questions about the level of adoption of terms and types, the shape and characteristics of entity descriptions and the distribution of data across the Web, for example, in terms of Pay Level Domains (PLDs), Top Level Domains (TLDs) or data publishers. In the following section we discuss the research questions and the methodology, followed by the data analysis and results of our study in the subsequent sections.

2 Methodology

2.1 Research Questions

The main target of this work is to answer certain questions regarding the usage of markup on scholarly data through a quantitative analysis. The research questions addressed in the following sections are:

- ***RQ1****: What are frequently used types and terms for scholarly data?* The main aim is to shape a better understanding of the adoption of vocabulary terms to comprehend the knowledge embedded through markup statements.
- ***RQ2****: How are statements about bibliographic data distributed across the Web and who are the key providers of bibliographic markup?* With this research question, investigated in Sect. 4, we research the distribution of data across domains and the indicated publishing institutions. We also aim to get a better understanding of the topic distribution, i.e., whether or not a strong bias towards particular scientific disciplines can be observed.
- ***RQ3****: What frequent errors can be observed?* In this context, we look into schema violations; significant syntactic and semantic errors introduced by data providers (Sect. 5).

[4] http://schema.org.
[5] http://www.webdatacommons.org.

These questions are approached through a quantitative analysis using the dataset described in the following section.

2.2 Methodology and Dataset

For our investigation, we use the Web Data Commons (WDC) dataset, being the largest available corpus of markup, extracted from the Common Crawl. Of the crawled web pages, 30% contain structured data which covers 17% pay-level-domains (PLDs)[6]. In addition, 20.48 billion RDF quads have been extracted, a significant amount when compared to DBpedia (4.58 million entities[7]) and Freebase (2.4 billion facts[8]). For our work we considered all statements which describe entities (subjects) that are of type *s:ScholarlyArticle* or of any type but co-occurring on the same document with any *s:ScholarlyArticle* instance.

To extract this subset, we processed the entire WDC2014 dataset using a Hadoop cluster for processing and extracting the investigated subset Our extracted dataset contains 6,793,764 quads, 1,184,623 entities, 83 distinct classes, and 429 distinct predicates. Due to space constraints, in later sections we will refer to *s:ScholarlyArticle* as *s:SchoArt*.

In our study, we have focused on schema.org as it is the most widely used schema and concentrated on *s:SchoArt*, *s:Person* and *s:Organization* for our analysis. Although there is a wide variety of types used for bibliographic and scholarly information, *s:SchoArt* is the only type which explicitly refers to scholarly articles. While this restricts our study with respect to recall, we followed this approach to enable a high precision of the analysed data within the scope of our study, where the goal is to provide conclusive insights into scholarly works markup only.

In order to identify related metadata to scholarly articles, our target was to find additional statements which relate the extracted instances of *s:SchoArt*. Since links between markup entities are sparse, the assumption that a node representing an author or affiliation of a specific article would be linked by the respective article instance does not hold true in the majority of cases. For this reason, we also consider instances of *s:Person* and *s:Organization* which co-occur with scholarly articles assuming that these will provide information about publishers or authors of the corresponding article.

3 Adoption of Scholarly Types and Predicates

This section addresses RQ1; we present an overview of utilized types and predicates in our extracted dataset. The major types considered are scholarly article (*s:SchoArt*), person (*s:Person*), and organization (*s:Organization*). Out of 6,793,764 triples and 1,184,623 entities, 280,616 instances are of *s:SchoArt*, 847,417 instances are of *s:Person* and 3,798 instances are of *s:Organization*.

[6] http://webdatacommons.org/structureddata/2014-12/stats/stats.html.
[7] https://en.wikipedia.org/wiki/DBpedia.
[8] https://en.wikipedia.org/wiki/Freebase.

Table 1. *Top-10* predicates used for *s:SchoArt*

Predicates	Occurrence
s:author	913,884
s:genre	204,954
s:image	191,879
s:headline	134,742
s:description	121,168
s:datePublished	119,448
s:publisher	115,896
s:keywords	104,488
s:name	78,873
s:editor	78,781

Table 2. *Top-10* PLDs according to the number of entities.

PLD	Entities	Statements
springer.com	850,697	3,011,702
bmj.com	106,777	877,589
mdpi.com	85,276	322,569
diabetesjournals.org	80,911	250,804
mendeley.com	42,564	143,376
biodiversitylibrary.org	24,946	122,457
gradesaver.com	24,108	121,592
grupoescolar.com	16,838	104,701
eurecom.fr	8,820	40,349
econjwatch.org	6,817	32,434

Among the organizations 1 instance is tagged as *s:Educational* organizations and 32 instances are tagged as *s:school* which is a further subtype of educational organization. Note that all the types and their subtypes are found by explicitly looking at the predicates for that particular type or subtype. For example, we have only captured those instances as *s:SchoArt* where the predicates corresponding to the instances specify scholarly article.

In Table 1, we present the *top-10* predicates ranked according to their occurrence. We find that for the type *s:SchoArt*, the predicate *s:author* depicts the highest occurrence with a frequency of 913,884. We also computed the number of distinct predicates for each instance for every extracted type. Figure 1 shows the distribution of distinct predicates over all the instances of the extracted types (*s:SchoArt*, *s:Person* and *s:Organization*). The number of distinct predicates for *s:SchoArt* varies from 1 to 14, for *s:Person* it from 1 to 9, and for *s:Organization* from 1 to 4.

It can be observed from both the distribution and the top-k predicates table, particular predicates are used very frequently, while there is a long tail of predicates which are hardly used. This provides insights as to the kind of knowledge which can be extracted from embedded scholarly data, where popular metadata is described in a fairly complete manner, for instance, author names and publication titles, while more specific information, for instance, about genres or publishers are less frequently found.

4 Distribution Across Domains and Documents

This section addresses RQ2, investigating the origins of bibliographic data, by analyzing the distribution of bibliographic markup across Pay-Level-Domains (PLDs), Top-Level-Domains (TLDs) and documents. There are 213 distinct PLDs, 38 TLDs and 199,979 documents across the subset.

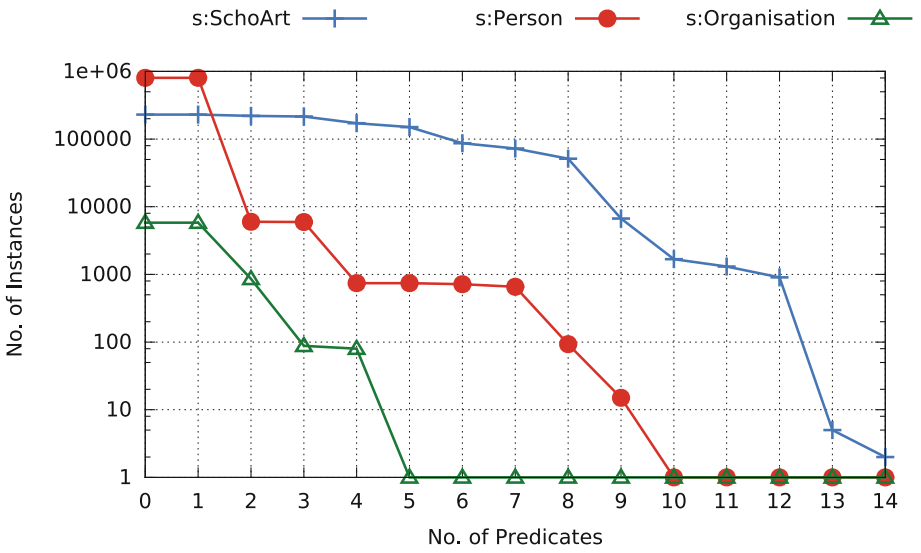


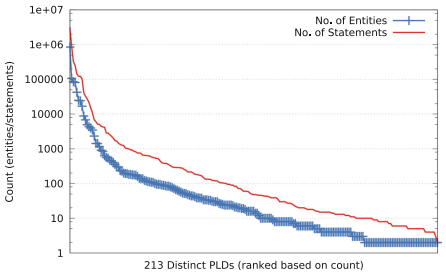
Fig. 1. Cumulative distribution of predicates over instances across extracted types. The number of instances (y-axis) are presented in log scale.

4.1 Distribution Across PLDs, TLDs and Documents

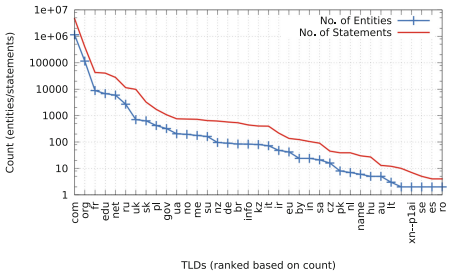
The distribution across domains and documents is represented in the plots of Fig. 2, where the blue (lower) line corresponds to the distribution of entities and the red (upper) line corresponds to the distribution of statements over PLDs, TLDs, and documents. The number of entities/statements presented on the *y-axis* are plotted in the logarithmic scale. As observed from the dataset, the number of statements is much higher than the number of entities corresponding to each PLD, TLD, or document. Another observation is the power law-like distribution of embedded markup across PLDs, TLDs, and documents, where

Table 3. Top-10 documents ranked according to embedded entities.

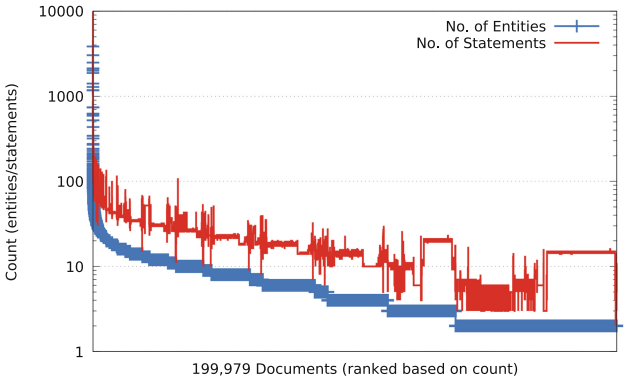
URL	Entities	Statements
<http://link.springer.com/article/10.1140%2Fepjc%2Fs10052-012-2183-y>	3843	7700
<http://link.springer.com/article/10.1007%2FJHEP02%282010%29041>	3035	6077
<http://www.russki-mat.net/page.php?l=FrFr\&a=C>	2486	9942
<http://link.springer.com/article/10.1140/epjc/s10052-010-1339-x>	2118	4242
<http://link.springer.com/article/10.1140/epjc/s10052-009-1227-4>	2114	4234
<http://link.springer.com/article/10.1140/epjc/s10052-010-1350-2>	2114	4234
<http://link.springer.com/article/10.1140%2Fepjc%2Fs10052-012-2175-y>	1999	4012
<http://www.chapman.edu/our-faculty/vernon-smith>	1879	5636
<http://cns.slis.indiana.edu/publications/>	1410	3507
<http://www.russki-mat.net/page.php?l=FrFr\&a=L>	1287	5144



(a) Distribution across PLDs



(b) Distribution across TLDs



(c) Distribution across HTML documents

Fig. 2. Distribution of entities/statements over PLDs, TLDs and documents. (Color figure online)

usually a small amount of sources provide the majority of entities and statements.

In Fig. 2(a) we plot the different PLDs along *x-axis* and the number of entities/statements corresponding to each PLD along *y-axis* in the logarithmic scale. We represent the PLDs in the ranked order of the number of entities and statements corresponding to them. For example, springer.com exposes a total of 850,697 entities and 3,011,702 statements. A detailed list of the *top-10* PLDs is shown in Table 2.

In Fig. 2(b) we plot the different TLDs along the *x-axis* and the number of entities/statements corresponding to each TLD along the *y-axis* in the logarithmic scale. For example, documents from *.com* domains expose 1,139,436 entities and 4,640,718 statements. As can be observed, *.com* and *.net* URLs are very frequent, while some national TLDs such as *.fr* appear among the top providers of scholarly bibliographic data. Basing our study on the assumption that the Common Crawl is a representative Web crawl, this provides first insights into the early adopters of embedded scholarly markup. A deeper look into the *top-k* PLDs supports the assumption that French academic and library institutions

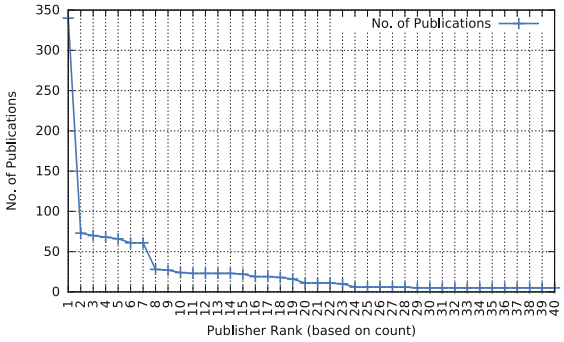


Fig. 3. Distribution of scholarly articles across publishers.

seem to be among the top providers of embedded markup. Similarly, Fig. 2(c) shows the distribution across HTML documents.

Tables 2 and 3 provide some insights into the most frequent PLDs (TLDs) and the documents including the highest amount of embedded data. We note that springer.com and *.com* are leading the queue in case of PLDs and TLDs respectively. On inspecting *top-10* PLDs, we observe that journals from the life sciences field, such as diabetesjournals.org and biodiversitylibrary.org are among the key data providers. This notion of a topic bias towards life and medicals sciences is further investigated in the following subsection.

On closer inspection, the documents which provide a significant amount of entities (top-10) often refer to pages about comprehensive publications, such as a book publication with rich annotation of bibliographic data, such as references for each chapter, as in the case of <http://link.springer.com/article/10.1140%2Fepjc%2Fs10052-012-2183-y> with 3843 embedded entities. Note that in rare cases (for instance, <http://www.russki-mat.net/page.php?l=FrFr&a=C> in row 3, referring to a Russian slang dictionary), flawed data is included, where instances are incorrectly typed and are actually not referring to scholarly data. This calls for further investigation into the correctness of embedded data (also see Sect. 5).

4.2 Distribution Across Topics and Publication Types

In order to better understand the topic coverage of scholarly data, we provide some initial insights into the most frequent publishers of detected scholarly articles, as indicated by the data itself, and the suspected topic bias of articles themselves. In Fig. 3 and Table 4 we show the overall distribution of scholarly articles across different publishers (533 distinct publishers in total) and the top-10 publishers ranked according to their publication count.

Similar to the distribution across PLDs, TLDs, and documents, the spread across publishers follows a power law distribution.

As observed in the table, most publishers seem to be either from the Computer Science domain (IEEE, Telecom Paris) or seem to be cross-domain, with a

Table 4. Top-10 publishers and their publication counts.

SchoArt:Publisher	#Publication
Econ Journal Watch	340
IEEE@fr	73
IEEE@en	70
TELECOM ParisTech@fr	68
TELECOM ParisTech@en	66
ENST Paris@en	61
ENST Paris@fr	61
Universit de Nice@fr	28
Universit de Nice@en	27
Springer@fr	24

Table 5. Most frequent publication types across the WDC dataset

SchoArt:genre	Article count
Article@en	7,788
Thesis@fr+@en	373
Conference@en+@fr	188
Journal@fr+@en	115
Rapport@fr+@en	16
Ouvrage@fr	7
Poster@en+@fr	8
Book@en	5
Talk@fr+@en	6
HDR@en+@fr	2
Others	6

Table 6. Top-10 article titles (pre-cleaned) ranked according to their occurrence.

Article title (SchoArt:name)	Occurrence
Highlights From the Latest in Diabetes Research@en	39
Essential information about patterns of victimisation among children with disabilities@en	36
Whose Oxis Being Gored? When Attitudinalism Meets Federalism@en	36
People with unhealthy lifestyle behaviours benefit from remote coaching via mobile technology@en	27
Longer duration of exclusive breastfeeding associated with reduced risk of childhood asthma up to age six@en	25
People with diabetes and selfreported severe hypoglycaemia have increased mortality risk over years@en	25
Community based nonpharmacological interventions delivered by family caregivers reduce behavioural and psychological symptoms of dementia@en	24
Preoperative physical therapy reduces risk of postoperative at elect as is and pneumonia in people undergoing elective cardiac surgery@en	24
How to Choose the Least Unconstitutional Option:Lessons for the President(and Others)from the Debt Ceiling Standoff@en	24
Post menopausal women with medically treated diabetes have increased risk of lung cancer@en	22

particular bias towards Life Sciences related literature (e.g. Springer). In order to get a clearer understanding of the actual topics of articles, we inspected the titles (*s:name*, *s:headline*) of articles. Although titles are often not well-populated we investigated frequently occurring titles, and ignored obviously noisy or misleading annotations.

From Table 6 we note that the top-10 actual article titles are all from the biological or medical domain, further indicating a strong inclination towards the Life Sciences.

In addition, we investigated the genre (*s:genre*) of detected articles, meant to describe the publication type. In Table 5, we cluster the genres such as thesis and journals having *@en* and *@fr* tags together to enhance readability. While articles (*Article(@en)*) seem by far the most used genre annotations, the whole range of bibliographic types is covered. Observed language annotations again confirm some bias towards English and French content and data providers.

5 Frequent Errors: Schema Violations

Errors are frequently found in embedded annotations, and the extent varies depending on the type of error. For instance, the use of undefined types and predicates is more frequent in traditional Linked Data, due to the fact that errors propagate through a dataset, as opposed to embedded data [4]. Other error types, such as schema violations and misuse of object properties are particularly frequent in embedded data. In Table 7, we report the most frequently misused predicates, that is predicates which are defined as object property but refer to data type/literal or vice versa. Here *S* and *P* are used as to indicate the range of the property, either <http://schema.org/ScholarlyArticle/> or <http://schema.org/Person/> respectively. For example *S:author* is an object property

Table 7. Top 10 misused predicates. Range refers to the expected range according to the schema.org vocabulary definition and is either *OP*-Object Property or *DP*-Data Type Property

Predicates	Occurrence	Range	Object	Data type	%Error
S:publisher	144147	OP	997	143150	99.31
S:creator	44615	OP	28550	16065	36.01
S:author	1048110	OP	697024	351086	33.49
S:about	888	DP	97	791	10.92
P:dateModified	7644	DP	419	7225	5.48
S:sourceOrganization	1637	OP	17	1620	1.01
P:affiliation	2144	OP	2129	15	0.69
S:headline	145953	DP	413	145540	0.28
S:datePublished	127494	DP	76	127418	0.06
P:editor	78781	OP	78773	8	0.01

having 1,048,110 occurrences within the dataset, where 697,024 instances correctly refer to a node (object), while the remaining 351,086 instances use it as a datatype property, directly referring to a literal (error rate 33.49%). From the Table 7 we can also observe that most often object properties are violated, while data type properties are largely compliant. This observation, further highlighting the lack of explicit links (object references) between entity descriptions in embedded markup, suggests that further research into coreference resolution and entity interlinking is required, in order to utilize scholarly markups as a potential knowledge graph.

6 Conclusion and Future Work

In this work, we have provided a first study on the coverage and characteristics of bibliographic metadata embedded in Web pages. Insights are provided with respect to frequent data providers, the adoption and usage of terms and the distribution across providers, domains and topics. The distribution in all cases follows a power law, with few providers and documents contributing the majority of data. The same can be identified for vocabulary terms, where few predicates are highly used, complemented by a long tail of predicates which are only used to a very small extent. With regard to the distribution across domains, a certain bias towards French data providers is observed based on manual investigation of the top-k genres and publishers. Article titles, PLDs, and publishers suggest a bias towards specific disciplines, namely Computer Science and the Life Sciences. However, the question as to what extent this is due to the selective content of the Common Crawl or representative for schema.org annotations on the Web in general, requires additional investigation.

As a part of future work, we are planning to conduct a follow-up study using a targeted crawl of typical providers of scholarly data (publishers, academic organizations, libraries), which would enable a more exhaustive and representative analysis. By limiting ourselves to explicitly annotated scholarly articles, it is also worth highlighting that a significant amount of bibliographic data has been excluded from our study. Here, as part of future work, other methods should be taken into account to classify implicitly typed bibliographic or creative work into scholarly or non-scholarly works. In addition, resolution of co-references and research into specifically tailored entity interlinking mechanisms would help to provide a more consolidated picture of the scholarly knowledge graphs which can be extracted from embedded data. This is an area where we see some key opportunities for related future work. Extracting (scholarly) knowledge graphs from Web documents provides opportunities for generating data far beyond the scale and dynamics of traditional datasets in the area. At the same time, embedded (scholarly) data can provide invaluable training data for targeted, i.e. domain-specific information extraction and linking algorithms for scholarly information.

References

1. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M., Völker, J.: Deployment of RDFa, microdata, and microformats on the web – a quantitative analysis. In: Alani, H., et al. (eds.) ISWC 2013. LNCS, vol. 8219, pp. 17–32. Springer, Heidelberg (2013). doi:10.1007/978-3-642-41338-4_2
2. Dietze, S., Taibi, D., dAquin, M.: Facilitating scientometrics in learning analytics and educational data mining the LAK dataset. Seman. Web J. (2015)
3. Meusel, R., Petrovski, P., Bizer, C.: The webdatacommons microdata, RDFa and microformat dataset series. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 277–292. Springer, Cham (2014). doi:10.1007/978-3-319-11964-9_18
4. Meusel, R., Paulheim, H.: Heuristics for fixing common errors in deployed *schema.org* microdata. In: Gandon, F., Sabou, M., Sack, H., d'Amato, C., Cudré-Mauroux, P., Zimmermann, A. (eds.) ESWC 2015. LNCS, vol. 9088, pp. 152–168. Springer, Cham (2015). doi:10.1007/978-3-319-18818-8_10