Journal of Web Engineering, Vol. 17, No.1&2 (2018) 045-072


# A QUANTITATIVE ANALYSIS OF THE USE OF MICRODATA FOR SEMANTIC ANNOTATIONS ON EDUCATIONAL RESOURCES

ROSA NAVARRETE

*Escuela Politécnica Nacional, Ecuador*
*rosa.navarrete@epn.edu.ec*

SERGIO LUJÁN-MORA

*University of Alicante, Spain*
*sergio.lujan@ua.es*

Received February 22, 2017
Revised September 24, 2017

A current trend in the semantic web is the use of embedded markup formats aimed to semantically enrich web content by making it more understandable to search engines and other applications. The deployment of Microdata as a markup format has increased thanks to the widespread of a controlled vocabulary provided by Schema.org. Recently, a set of properties from the Learning Resource Metadata Initiative (LRMI) specification, which describes educational resources, was adopted by Schema.org. These properties, in addition to those related to accessibility and the license of resources included in Schema.org, would enable search engines to provide more relevant results in searching for educational resources for all users, including users with disabilities. In order to obtain a reliable evaluation of the use of Microdata properties related to the LRMI specification, accessibility, and the license of resources, this research conducted a quantitative analysis of the deployment of these properties in large-scale web corpora covering two consecutive years. The corpora contain hundreds of millions of web pages. The results further our understanding of this deployment in addition to highlighting the pending issues and challenges concerning the use of such properties.

*Keywords*: semantic web, Microdata, educational resources, Schema.org, LRMI, educational resources, web standards

*Communicated by*: M. Gaedke & L. Olsina

## 1 Introduction

The increasing interest in using the semantic web for exploring information based on its meaning has promoted the use of structured markup formats to add semantic annotations to web content. These semantic annotations are integrated into human-readable web pages to make the content machine-processable and unambiguous, so that search engines can enhance the accuracy and visualization of search results. The most common formats used for semantic annotations are Microdata, Microformats, JSON-LD and RDFa [1]. Of these formats, Microdata has been the most broadly adopted [2] and has

spread across a wide range of topics, such as personal information, events, products, organizations, movies, recipes and more [3].

Meanwhile, there is a growing trend in the availability of educational resources on the Web [4]. These resources can be tagged with descriptors of different metadata standards for their identification and retrieval [5] and they are usually stored in repositories available through websites. This raises the problem of finding relevant educational resources to match the specific needs of educators and learners, such as subject context, educational level, language, format, pedagogical approach and the associated license (i.e., the conditions of use of the resource). The ability to search for and retrieve educational resources on the Web is essential for the efficient use of such resources [6, 7] and this is a critical issue for people with disabilities because of their requirements regarding web accessibility [8].

Finding accessible educational resources for people with disabilities is relevant in the current context of education and society, where the Web is the prevailing medium for communication and continuous learning [9]. At present, according to the World Health Organization [10], people with disabilities represent around 15% of the global population and this share is increasing because of the disabilities caused by an aging population. The United Nations [11] has projected that, globally, by 2030, the number of individuals aged over 60 will have grown by 56% to 1.4 billion, and will reach nearly 2.1 billion by 2050.

In this context, the use of Microdata for the embedded markup of educational content can provide meaningful information that enables search engines to parse semantic content to improve the indexing, searching, and retrieval of resources [12]. Consequently, users could benefit from relevant search results based on their needs [13] regarding the educational value of the resources, their accessibility characteristics, and their license of use.

The most widespread standard for Microdata vocabulary is Schema.org [14]. This specification sets is a collection of shared vocabularies that include properties to describe characteristics of broad types of web content. These descriptions can be understood by major search engines [15]. In 2013, Schema.org adopted the Learning Resource Metadata Initiative (LRMI) specification [16], which is a collection of properties used to describe educational resources on the Web. The properties adopted from LRMI can ideally be used in conjunction with other properties that belong to Schema.org to form a complete description of an educational resource. These complementary properties are intended to describe accessibility characteristics and the resource license.

In this work, we investigated the spread of the use of embedded markup with Microdata to improve educational resource web searches. We analyzed the co-occurrence of the deployment of three sets of properties from Schema.org that are considered relevant when searching for educational resources. These sets of properties, as explained above, are related to educational value (properties from LRMI), accessibility characteristics, and the resources' license of use. We assessed whether these three sets of properties are used conjunctly to describe a resource through a procedure that examines the embedded markup on each web page where these properties appear. This diagnostic delivers a quantitative analysis of the degree of deployment of Microdata used to describe educational resources.

This quantitative analysis was conducted on datasets extracted from the Common Crawl Corpus [17], as it is the largest corpus of web crawl. The datasets containing structured data were extracted by the Web Data Commons (WDC) project [18] and are available for public use. Two datasets were

considered: the first, from December 2014, with 2.01 billion pages, of which 620 million pages correspond to structured data; and the second, from November 2015, with 1.77 billion pages, of which 541 million pages correspond to structured data.

There are few previous studies that focus on the use of Microdata in the educational field from a quantitative perspective. Analysis of the deployment of a group of properties related to the educational field is addressed in [19] and is related to the scholarly field in [20]. In this research, we extended the analysis by verifying the deployment of Microdata with three sets of properties from Schema.org in educational resources on the web. Since this research is based on credible and recent large-size web corpora from December 2014 and November 2015, it is possible to obtain a first reliable insight into the trend of adoption in the deployment of Microdata for education.

Moreover, because each dataset was processed, the outcomes of the procedure applied in this research include data regarding all web domains that use properties belonging to the sets analyzed, not only those domains related to the educational field. Furthermore, issues related to the inappropriate use of properties under analysis are reported.

The outcomes of this research are intended to spread awareness about the potential use of Microdata annotations on educational resources to encourage their adoption by web developers, authors of educational web content, and developers of web content authoring tools. Moreover, the results could be valuable to the W3C Schema.org Community Group because identifying the actual extent of use of the Schema.org standard in the educational field could be considered useful for future decisions about the versioning of this standard.

The structure of this paper is as follows: section two presents a review of the standards of Microdata; section three outlines a systematic literature review to present related works; section four describes the corpora basis for analysis, as well as the parameters, research questions, and the applied procedure; section five presents and discusses the main findings; and, finally, section six presents conclusions about the results and discusses any detected problems.

## 2 Microdata and Schema.org

The Microdata format has been developed in the context of the standardization of HTML5 for the embedded markup of web content [21]. Microdata consists of a group of name-value pairs; the groups are called items and each name-value pair is a property. Items are defined using the following five attributes: itemscope; itemtype; itemid; itemprop, and itemref [22]. These annotations provide the semantics through the terminology and properties of a knowledge representation domain including the inherent relationships [1, 7]. By way of example, Figure 1 presents a part of the HTML5 code with Microdata annotations taken from a web page of Bookshare (www.bookshare.org), an accessible online library that provides books for people with visual impairments. In this example, Microdata properties are used to describe an educational resource, in this case a textbook, by using the syntax:

itemprop = “property-name” content = “property-value”

Several properties analyzed in this study are included in the sample of Figure 1; for instance, a description of accessibility characteristics (*accessibilityFeature*), a description of information about the license (*license*), an indication of the recommended age of resource users (*typicalAgeRange*), a

definition if an interactivity type is expected (*interactivityType*), and a definition of the type of resource (*learningResourceType*).

```
<meta itemprop="bookFormat" content="EBook/DAISY3"/>
<meta itemprop="accessibilityFeature" content="displayTransformability/font-size"/>
<meta itemprop="accessibilityFeature" content="displayTransformability/font-family"/>
<meta itemprop="accessibilityFeature" content="displayTransformability/color"/>
<meta itemprop="accessibilityFeature" content="displayTransformability/background-color"/>
<meta itemprop="accessibilityFeature" content="bookmarks"/>
<meta itemprop="accessibilityFeature" content="readingOrder"/>
<meta itemprop="accessibilityFeature" content="structuralNavigation"/>
<meta itemprop="accessibilityHazard" content="noFlashingHazard"/>
<meta itemprop="accessibilityHazard" content="noMotionSimulationHazard"/>
<meta itemprop="accessibilityHazard" content="noSoundHazard"/>
<meta itemprop="accessibilityControl" content="fullKeyboardControl"/>
<meta itemprop="accessibilityControl" content="fullMouseControl"/>
<meta itemprop="license" content="https://www.bookshare.org/cms/legal-information"/>
<meta itemprop="provider" content="Bookshare.org"/>
<meta itemprop="url" content="https://www.bookshare.org/browse/book/1056770"/>
<meta itemprop="typicalAgeRange" content="18-"/>
<meta itemprop="interactivityType" content="expositive"/>
<meta itemprop="learningResourceType" content="textbook"/>
<meta itemprop="audience" content="student"/>
```

**Figure 1.** Example of the use of Microdata embedded in HTML5 code

Further, embedding Microdata in HTML tags is a simple way to add semantic annotations and they can be parsed into RDF [23], providing a way to publish Linked Data [24] to interlink data into the Web.

Although annotations with Microdata can use arbitrary language for the "itemtype" and "itemprop" values, this raises the problem that search engines do not understand the meaning of the content. Therefore, Microdata annotations can be enhanced with the use of a controlled vocabulary such as the one provided by Schema.org through a set of classes and properties [15]. These annotations are easily understandable by users and simple for search engines to process in order to search the content effectively [25]. Search engines can use these annotations to enrich search results [13, 26] with, for example, data snippets, data tables, and answers to fact questions.

Since its launch in 2011, the main search engine companies (Google, Microsoft Bing, Yahoo and Yandex) and the World Wide Web Consortium (W3C) have driven the Schema.org initiative. Due to this use, Schema.org has become the de facto standard for embedded markup [14, 27]. Schema.org has expanded to cover a broad range of topics. Initially, the standard was composed of 297 classes and 187 relationships; this has increased to 638 classes and 965 relationships [3]. All classes are organized into a hierarchy and relationships can be used for more than one topic.

In April 2013, Schema.org adopted the LRMI version 1.1 (2011) specification that is currently under development by the Dublin Core Metadata Initiative [16]. These properties are called "LRMI properties" for the rest of the paper and are intended to describe educational resources by adding specific properties to make them easily locatable through search engines and search services [28].

## 3 Related works

We carried out a literature review by applying a systematic concept-centric approach [29]. The results enabled us to both discover what studies have already been carried out in relation to this research and contextualize the contribution of this research.

### *3.1 Process for systematic literature review*

The review process included the following steps:

a. *Selection of the input source for searching research publications.* The search was conducted using four research literature databases: Web of Science (WOS); Scopus; IEEE Xplore; and ACM Digital Library (ACM DL).

b. *Selection of search criteria.* The criteria defined for the search were related to the search string, the search place within the paper, the type of publication, the year of publication, and the language used.

   The search string used for the search was:

   ("Microdata" AND ("semantic web" OR "Schema.org" OR"semantic annotations"))
   OR ("Schema.org" and "markup") OR ("corpus web" and "structured data")

   This search string prevented ambiguity of the term "Microdata" (also used in the statistical field) and ensured that all fields involved in the scope of this paper were included.

   The search string was applied for searches on title, abstract and keywords. Only journal and conferences articles were included. In addition, the years of publication were restricted to 2011 onwards because the search engines Google, Bing, Yahoo and Yandex created Schema.org in 2011 [3]. Finally, only papers written in English were considered.

   There was the potential for the search results for the research literature databases to overlap. The same article could be reported as a search result in more than one of the sources. Thus, articles were considered only once.

c. *Selection of applicability topics.* The articles found in the search were reviewed in order to define their applicability to this study. The topics considered as relevant for this work were the following:

   - The use of Microdata for enhancing web searches;
   - Qualitative or quantitative analysis of Microdata applications in specific contexts;
   - The use of the Schema.org vocabulary; and,
   - Web corpora and analysis of the use of structured data.

   In this step, the backward technique was used to extend the literature by recovering other articles included in the reference list of papers under review.

d. *Synthesize the literature*. The final step entailed linking the literature under review to the proposed work.

### 3.2 *Results of literature review*

The results of the systematic literature review are presented by displaying the number of research articles found in search databases, the synthesis of the articles according to applicability topics previously defined, and the contribution of this work.

#### *3.2.1 Number of research articles found in search databases*

The results of the systematic literature review are presented in Table 1. The columns display the number of articles found in each research literature database by year and the total research articles reviewed based on the applicability topics. The additional articles found by using the backward technique are also included. If the research article appeared in more than a source, it is counted in the oldest source according to its publication date. The bottom of Table 1 presents the total number of research articles considered as works related to this proposed research.

**Table 1.** Results of literature review

| **Source** | **2011** | **2012** | **2013** | **2014** | **2015** | **2016** | **Total** |
|---|---|---|---|---|---|---|---|
| Scopus | 0 | 1 | 6 | 3 | 3 | 2 | 15 |
| WOS | 1 | 7 | 14 | 7 | 6 | 5 | 40 |
| IEEE Xplore | | 3 | 4 | | 2 | | 9 |
| ACM DL | | | 1 | 2 | | 3 | 6 |
| Total research articles | 1 | 11 | 25 | 12 | 11 | 10 | 70 |
| Relevant research articles | 1 | 2 | 4 | 3 | 1 | 3 | 14 |
| Additional research articles (backward technique) | 1 | 3 | 2 | 2 | 1 | 2 | 11 |
| **Total research articles as related works** | **2** | **5** | **6** | **5** | **2** | **5** | **25** |

#### *3.2.2 Synthesis of articles according to applicability topics*

To synthesize the related works, each article is associated with the applicability topics previously defined, as set out below. Some articles are associated with more than one topic.

*a. Use of Microdata for enhancing web searches*

These works support the spread of Microdata as a strategy to generate semantic annotations that enhance web searches.

In [13, 30], markup formats such as Microdata are proposed to enhance web search results and, consequently, users' experience. The use of all markup formats including Microdata is analyzed in [31] as a mechanism to tag web content with machine-readable information. In [32], new Microdata vocabulary is proposed in conjunction with a semi-automatic semantic annotation method to improve the structured Web of Things resources.

In [33], the authors propose the use of semantic technologies, such as the implementation of Microdata, to add semantic markup to HTML content in a university digital repository. In [34], the authors propose the inclusion of new Microdata schemas for describing 3D media objects aimed at enhancing the search results of this content. These new properties are related to content, spatial temporal, structural, logical, and behavioral features.

*b. Quantitative or qualitative analysis of Microdata applications in specific contexts*

These works address the analysis of the use of Microdata in different contexts from either a quantitative or qualitative perspective.

Regarding the quantitative approach, we found two works that focus on the use of Microdata in the educational context. These works analyze different sets of properties: [19] examines the use of properties based on the LRMI specification and [20] examines the use of properties to describe scholarly information. Both studies are carried out on WDC datasets.

Moreover, [35] studies the use of structured data in e-commerce. This study includes all markup formats, including Microdata, and is conducted over the one million most popular online shop websites.

On the other hand, there are several qualitative studies that focus on the use of Microdata. In [36], the author addresses the use of Microdata in embedded markup for videos in order to describe the properties for searching, archiving, and processing this type of media. The use of Microdata in Russian video content delivery sites is presented in [37].

In [12], the author proposes the use of Microdata as a new approach for describing educational resources. Another study [38] reviews the use of Microdata for semantic searches in wikis. In [39], the author suggests the use of Microdata for marking up web content as a strategy for achieving an educational semantic web.

In [40], the author presents the use of semantic annotations through Microdata in web shops. Another study [41] proposes Microdata annotations to describe products in a catalog for e-commerce at web scale. A different study [42] addresses several topics related to the web of data for e-commerce, the adoption of the GoodRelations vocabulary, and the official e-commerce model of Schema.org as well as its implementation by using markup formats such as Microdata. In [43], the authors propose that automatic annotations with markup formats, such as Microdata, should be used to enhance product advertisements. This strategy makes it possible for a large amount of product description data to be made publicly available.

*c. Use of Schema.org vocabulary*

These works address the evolution of Schema.org vocabulary by proposing new terms and relationships to be added to this vocabulary. In [25], Schema.org is used to reduce ambiguity on the web pages of open digital libraries. In [28], the use of Schema.org vocabulary is extended by using LRMI properties to describe educational resources. In [44], the mapping of Schema.org with Linked Data is presented using an empirical approach. In [45], the authors propose an extension of Schema.org vocabulary to describe web observatories (i.e., platforms to collect data about the Web).

*d. Web corpora and analysis of use of structured data*

These articles relate to the WDC project and present the extraction of web corpora, as well as the analysis of the deployment of markup formats. In [2], the authors present the extraction of datasets of structured data from the Common Crawl Corpus and statistics about the markup formats within these datasets. Another study [46] presents, the extraction process from the Bing Crawler corpus in comparison with the extraction of the WDC corpus. This study also includes statistics about the use of

markup formats. Two studies [47, 48] conduct a statistical analysis of the use of different markup formats (Microdata, Microformats, RDFa) in the WDC corpus.

*3.2.2 Contribution of this work*

From the results of the systematic review, we can argue that there are no previous studies that address the use of Microdata aimed to enhance the search for educational resources. This study enlarges the work [19] by focusing on the analysis of the use of three sets of properties aimed to describe educational resources. We obtained detailed information about domains that use these annotations, the frequency of use of each property, and the issues related to the misuse application of properties.

Furthermore, to visualize a trend in the deployment of this set of properties, we analyzed web corpora of the two consecutive years following the adoption of LRMI in 2013.

## 4 Basis for analysis

This section describes the web corpora under analysis, the sets of Microdata properties that were analyzed, the research questions, and the procedure used for analysis.

### 4.1 *Web corpora for analysis*

The Common Crawl Foundation (CCF) extracts data from the Web based on a snapshot of the most popular part of the Web defined by the PageRank [17]. This extraction process retrieves web pages from pay-level domains (PLDs), i.e., any subdomain of a public top-level domain (TLD), which allows us to assert that this domain is in the control of an organization or a user [49].

The CCF periodically releases the web corpus results from each crawling process for public use. This is the Common Crawl Corpus. The websites and the number of web pages from each website vary in each crawling process. The datasets for analysis are extracted from the large-scale Common Crawl Corpus by the WDC project [18]. This project extracts from the Common Crawl Corpus minor subsets of data containing information on web pages that use structured data and release these for public use.

For this work, we considered the datasets released in December 2014 and November 2015 that contain web pages that use structured data. These subsets of data were the corpora for this analysis and are named “Corpus 2014” of structured data (Microdata, Microformats, and RDFa) [50] and “Corpus 2015” of structured data (Microdata, Microformats, JSON-LD, and RDFa) [51], respectively. The format JSON-LD was incorporated in Corpus 2015.

The datasets are delivered as files of plain text that contain n-quads that represent a set of RDF data. Each n-quad is written in a single line formed by a sequence of terms of a triple RDF (subject, predicate, object) and a label of a blank node or an IRI label (could be URI or URL) to identify the set of data from which the triple has been extracted [2, 52].

Information about the Common Crawl Corpus and Microdata datasets that were analyzed in this work are presented in Table 2. The column named Code in Table 2 is used to refer to respective data in the rest of this paper.

Considering the Common Crawl Corpus of both releases, Corpus 2015 has a lower number of web pages than Corpus 2014 (see CC1 in Table 2). Further, the percentage of web pages containing

structured data is very similar in both corpora (WDC1 in Table 2). Nevertheless, the number of web pages with Microdata annotations as well as the domains that use such annotations, increases significantly from Corpus 2014 to Corpus 2015 (MIC1 and MIC2 in Table 2).

**Table 2.** Information about corpora for analysis

| Code | Subject | Corpus December 2014 | | Corpus November 2015 | |
|---|---|---|---|---|---|
| **Common Crawl Corpus** | | | | | |
| CCC1 | Total web pages | 2,014,175,679 | | 1,770,525,212 | |
| CCC2 | Total domains | 15,668,667 | | 14,409,425 | |
| **WDC Corpus with structured data (Microdata, Microformats, JSON-LD, and RDFa)** | | | | | |
| WDC1 | Web pages | 620,151,400 | 30.79% (of CC1) | 541,514,775 | 30.58% (of CC1) |
| WDC2 | Domains | 2,722,425 | 17.37% (of CC2) | 2,724,591 | 18.91% (of CC2) |
| **Microdata datasets analyzed** | | | | | |
| MIC1 | Web pages | 292,601,824 | 47.18% (of WDC1) | 312,229,919 | 57.66% (of WDC1) |
| MIC2 | Domains | 819,990 | 30.12% (of WDC2) | 1,100,783 | 40.40% (of WDC2) |
| MIC3 | Number of files of plain text | 2,301 | | 2,987 | |
| MIC5 | Size of data (decompressed) | 2,471,611,516,156 Bytes (2.24792 Terabytes) | | 3,512,571,049,622 Bytes (3.19466 Terabytes) | |
| MIC6 | Number of n-quads | 9,451,742,113 | | 13,514,697,971 | |

### 4.2 *Microdata properties*

The target set of properties for analysis includes LRMI, accessibility, and license properties. These are explained in the following three subsections.

#### 4.2.1 *LRMI properties*

Schema.org recently adopted this set of properties in 2013. Table 3 shows the LRMI properties from the LRMI specification within the respective classes of Schema.org in which they were included [16]. The LRMI properties included in Table 2 can be grouped into:

a. Educational LRMI properties specifically targeted towards educational aspects. These are the following: *educationalAlignment*, *educationalUse*, *learningResourceType*, *aligmentType*, *educationalFramework,* and *educationalRol*e.

b. General purpose LRMI properties. These are the following: *typicalAgeRange*, *timeRequired*, *interactivityType*, *isBasedOnUrl*, *targetDescription*, *targetName*, and *targetUrl*. These properties can also be used for any type of content, such as computer games.

The column named "Values" shows the allowed values of each attribute that are part of a controlled vocabulary or the patterns for possible values to exemplify their use. In other cases, some examples of possible values are presented. The admissible values for the *educationalFramework* property are contextual to educational systems in certain countries or regions.

**Table 3**. LRMI properties in Schema.org

| Property | Description | Values |
|---|---|---|
| **Class in Schema.org: CreativeWorks** | | |
| educationalAlignment | Defined on alignmentObject. | Text |
| educationalUse | The purpose recommended for the educational resource in the pedagogical context. | "assignment", "group work" |
| timeRequired | The average time for reviewing the resource. | "PT30M", "PT1H25M"[1] |
| typicalAgeRange | The typical age recommended for using the resource. | "12-18", "18-" |
| interactivityType | The learner's interaction with the resource. | "active", "expositive", "mixed" |
| learningResourceType | The predominant type of resource. | "presentation", "lecture", "lesson plan", "learning activity" |
| isBasedOnUrl | The URL of a resource that was used in the creation of this resource. | URL |
| **Class in Schema.org: AlignmentObject** | | |
| alignmentType | The category of alignment between the learning resource and the educational framework. | "assesses", "teaches", "requires", "textComplexity", "readingLevel", "educationalSubject", "educationLevel" |
| educationalFramework | The framework to which the resource being described is aligned. | Standard for Professional Engineering Competence[2], Professional Education and Training[3], Common Core State Standards[4] |
| targetDescription | The description of a node in an established educational framework. | Text |
| targetName | The name of a node in an established educational framework. | Text |
| targetUrl | The URL of a node in an established educational framework. | Text |
| **Class in Schema.org: EducationalAudience** | | |
| educationalRole | The role that describes the target audience of the content. | "learner", "teacher" |

[1]https://en.wikipedia.org/wiki/ISO_8601#Durations
[2]http://www.engc.org.uk/ukspec.aspx
[3]http://www.lawscot.org.uk/education-and-careers/
[4]http://www.corestandards.org/

Table 2 shows possible value for the *educationalFramework* property such as "Common Core State Standards" (the set of standards established in the US education system to ensure high-school completion) and "Standard for Professional Engineering Competence" (establishes the competencies required for a professional engineering registration in the UK).

### *4.2.2 Accessibility properties*

Accessibility properties are intended to describe how people with disabilities could process web content. For example, with the use of the *accessibilityFeature* property it is possible to include information about the type of transformability that enables the resource in terms of font size, background and foreground color, and reading order.

Table 4 shows the properties used to describe accessibility characteristics that are included in the CreativeWork class of Schema.org [15]. These properties are termed "accessibility properties" for the rest of this paper.

**Table 4.** Accessibility properties in Schema.org

| Property | Description | Values |
|---|---|---|
| accessibilityAPI | The compatibility of the resource with the referenced accessibility API. | "AndroidAccessibility", "ARIA", "iOSAccessibility", "JavaAccessibility" |
| accessibilityControl | The sufficient input methods to control the resource. | "fullKeyboardControl", "fullMouseControl", "fullSwitchControl", "fullTouchControl", "fullVideoControl", "fullVoiceControl" |
| accessibilityFeature | The content features of the resource, such as accessible media, alternatives and supported enhancements for accessibility. | "alternativeText", "audioDescription", "bookmarks", "braille", "captions", "ChemML", "displayTransformability", "highContrastDisplay", "longDescription", "readingOrder", "signLanguage", "structuralNavigation", "tableOfContents", "taggedPDF", "transcript" |
| accessibilityHazard | The characteristic of the resource that is physiologically dangerous to some users (flashing light, sound, and motion simulation). | "flashing", "noFlashingHazard", "motionSimulation", "noMotionSimulationHazard", "sound", "noSoundHazard" |

We also verified the use of accessibility properties of the pending extensions of Schema.org proposed by the Epub 3.1 Accessibility Working Group [53] (*accessMode*, *accessModeSufficient*, *accessibilitySummary*). Although the pending extensions contain attributes that have not yet been formally accepted by Schema.org, their potential application is relevant to the context of this analysis.

Further, accessibility properties are not exclusively linked to educational content. These properties can be used to describe other types of content, such as entertainment videos.

#### 4.2.3 *License properties*

We analyzed annotations concerning the type of resource license because it is relevant for users to know the restrictions of use of the resource; for example, this property can identify an open educational resource that can be used freely. License annotations can be done by using the property *license* or *useRightsUrl*. These properties are termed "license properties" for the rest of this paper.

The property named *license* is part of Schema.org, while the property named *useRightsUrl* belongs to the LRMI specification but was not adopted by Schema.org [54]. However, it is relevant to know whether developers follow the standard. Further, these properties can be used to describe the license of other types of content, such as publicity images.

### 4.3 *Research questions*

Analysis was conducted to address the following main research questions (RQ):

RQ1. How extensive is the deployment of each property belonging to the sets of LRMI properties, accessibility properties, and license properties? The results should be exposed separately by each set and contain the number of n-quads where the property appears.

RQ2. How extensive is the use of each property belonging to the sets of LRMI properties, accessibility properties, and license properties in relation to the domains that use them? The results should consider educational and non-educational domains.

RQ3. Have the properties of the sets of LRMI properties, accessibility properties, and license properties been used for the expected purpose?

RQ4. Have the properties of the sets of LRMI properties, accessibility properties, and license properties been used together to improve the description of educational resources?

It is important to note that the number of n-quads is not a reliable indicator of the extent of the adoption of Microdata or of the quality of use of the properties, because the domains that publish content in bulk increase the number of n-quads. Nevertheless, the contribution of n-quads still comes from just one domain. For this reason, RQ2 focuses on the results of the use of properties in relation to the domains that use such properties.

RQ1 and RQ2 offer a quantitative approach for analysis, while the analysis for RQ3 and RQ4 is qualitative. RQ1 and RQ2 can be presented together for an analysis of the use of properties that show n-quads in relation to domains.

### 4.4 *Procedure for analysis*

The Microdata datasets for analysis from Corpus 2014 and Corpus 2015 are those tagged with MIC3 in Table 2. The files that configure the datasets are indexed as a file with a ".list" extension that contains references to a set of compressed files with a ".gz" extension. We opted to download the files to local storage for processing.

The procedure for analysis required a compute-intensive task to process the files, one at a time, by using several processors. The process followed these steps:

- Decompress each file on the list;
- Open the file and examine each n-quad to extract its components. Each n-quad that includes the use of any of the properties belonging to the sets of properties proposed in this work was separated for analysis; and,
- Group n-quads by property and domain.

Although an analysis was conducted for all proposed sets of properties (i.e., LRMI properties, accessibility properties, and license properties), the results are provided separately by each set. Moreover, to address RQ2, we categorized the domains as educational and non-educational under the assumption that only educational domains release educational resources. To determine the educational domains encountered in the analysis, we considered the categorization of web domains defined by the DMOZ Internet Directory [54] and the manual reviews of each website by an expert for validation.

Educational resources can adopt different formats, such as a book, image, text document, web page, video, etc. Therefore, the properties of *Schema.org/CreativeWorks* can be used in conjunction with other classes such as *Book*, *WebPage*, *Image*, among others [55]. For this work, we considered the use of the property regardless of the class in which it is included by Schema.org.

As part of the process, n-quads with formatting errors were discarded from the statistics and were stored for manual verification. The file processing involved the extraction of data from files and the analysis of each element of n-quads. Some incorrectly formatted n-quads were detected in the process. For example, sometimes characters that are not allowed are included in the middle of the elements, such as escape characters, blanks in the middle of the URL, etc. These n-quads were not considered for evaluation because they could not be correctly processed.

## 5 Results and discussion

The results of the analysis are presented linked to the corresponding research questions applied to the LRMI properties (Table 3), accessibility properties (Table 4), and license properties (*license* and *useRightsUrl*), considering both educational and non-educational domains. In some cases, the presentation of the results of the quantitative analysis involves aspects of both RQ1 and RQ2.

### 5.1 *Analysis parameters*

#### 5.1.1 *Use of LRMI properties*

The results tables first list the educational LRMI properties, followed by the general purpose LRMI properties, and they are visually separated by a thick line in the table. To establish an order of presentation, the information is shown in descending order, according to the number of n-quads of each property in Corpus 2014.

Regarding RQ1 in terms of LRMI properties, Table 5 presents the number of n-quads of each LRMI property in the corpus and the percentage that this number represents in relation to the total number of n-quads in each corpus. In relation to RQ2, Table 5 presents the number of domains that use each property and the percentage that the property represents in relation to the total number of domains in each corpus. Each domain can use more than one property. Table 5 shows that the total number of n-quads using these properties in Corpus 2014 is 1,799,281, which represents 0.019% of the total number of n-quads with Microdata from this corpus (9,451,742,113 MIC6 of Corpus 2014 in Table 2). The total number of n-quads using these properties in Corpus 2015 is 2,232,159, which represents 0.016% of the total number of n-quads with Microdata from this corpus (13,514,697,971 MIC6 of Corpus 2015 in Table 1).

**Table 5.** N-quads and domains for LRMI properties

| Property | Corpus 2014 | | | | Corpus 2015 | | | |
|---|---|---|---|---|---|---|---|---|
| | n-quads | % Total n-quads | Domains | % Total domains | n-quads | % Total n-quads | Domains | % Total domains |
| educationalAlignment | 188,103 | 10.45 | 12 | 4.26 | 276,544 | 12.39 | 13 | 3.16 |
| alignmentType | 90,941 | 5.05 | 9 | 3.19 | 197,604 | 8.85 | 12 | 2.92 |
| learningResourceType | 82,758 | 4.60 | 36 | 12.77 | 126,664 | 5.67 | 67 | 16.30 |
| educationalUse | 51,488 | 2.86 | 22 | 7.80 | 60,275 | 2.70 | 44 | 10.71 |
| educationalRole | 8,543 | 0.47 | 2 | 0.71 | 6,972 | 0.31 | 4 | 0.97 |
| educationalFramework | 1 | 0.00 | 1 | 0.35 | 3,288 | 0.15 | 2 | 0.49 |
| isBasedOnUrl | 718,216 | 39.92 | 127 | 45.04 | 293,881 | 13.17 | 172 | 41.85 |
| typicalAgeRange | 302,304 | 16.80 | 94 | 33.33 | 550,212 | 24.65 | 124 | 30.17 |
| targetName | 137,323 | 7.63 | 3 | 1.06 | 120,223 | 5.39 | 6 | 1.46 |
| timeRequired | 94,748 | 5.27 | 44 | 15.60 | 225,777 | 10.11 | 58 | 14.11 |
| targetDescription | 97,061 | 5.39 | 4 | 1.42 | 86,142 | 3.86 | 5 | 1.22 |
| interactivityType | 26,795 | 1.49 | 25 | 8.87 | 284,364 | 12.74 | 49 | 11.92 |
| targetUrl | 0 | 0.00 | 0 | 0 | 213 | 0.01 | 1 | 0.24 |
| **Total in Corpus** | **1,799,281** | | **282** | | **2,232,159** | | **411** | |

The percentage of use of these properties in Corpus 2015 shows a slight decrease from Corpus 2014. The *targetUrl* property is not detected in Corpus 2014 and has a minimum use in Corpus 2015, while *educationalFramework* and *educationalRole* have a minimum use in both corpora. It should be noted that the percentage of use of *isBasedOnUrl* property decreases from 39.92% of the total number of n-quads in Corpus 2014 to 13.17% of the total number of n-quads in Corpus 2015. This is most likely because this property of LRMI is superseded by *isBasedOn* from Schema.org.

Regarding RQ2, we reviewed the properties by domain in each corpus. In Corpus 2014, we found that at least one of these properties can be detected in 282 domains, of which 12 are educational domains (4.26%) and 270 (95.74%) are for diverse purposes (personal blogs, advertising companies, bookshops, media companies). In Corpus 2015, we found that at least one of these properties could be detected in 411 domains, of which 28 domains (6.82%) correspond to the educational field and 383 (93.18%) are for diverse purposes (personal blogs, advertising companies, bookshops). The number of domains that use LRMI properties increases from 282 in Corpus 2014 to 411 in Corpus 2015 (45.74%); within these domains, the educational domains also increase from 12 in Corpus 2014 to 28 in Corpus 2015 (133.33%).

Furthermore, we expanded the results related to RQ1 and RQ2 by presenting the number of n-quads of each property in the educational and non-educational domains identified in both corpora, as explained below.

Table 6 presents educational domains in Corpus 2014 with the number of n-quads on each LRMI property. Table 7 presents the same information for Corpus 2015. In both tables, the columns of properties are presented in the same order as Table 5. Further, the comma to separate groups of thousands has been omitted in all columns of properties because of space constraints inside the table.

**Table 6.** Use of LRMI properties in educational domains in Corpus 2014

| Domain | n-quads | educational Alignment | aligmentType | learning ResourceType | educational Use | educational Role | isBasedOnUrl | typical AgeRange | targetName | time Required | target Description | interactiviy Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| www.brainpop.com | 550,219 | 97046 | 0 | 0 | 0 | 0 | 162035 | 97046 | 97046 | 0 | 97046 | 0 |
| www.merlot.org | 129,368 | 40276 | 40276 | 0 | 0 | 8540 | 0 | 0 | 40276 | 0 | 0 | 0 |
| phet.colorado.edu | 123,352 | 49726 | 49726 | 4780 | 14340 | 0 | 0 | 0 | 0 | 0 | 0 | 4780 |
| www.teacherspayteachers.com | 52,995 | 0 | 0 | 17665 | 0 | 0 | 0 | 17665 | 0 | 17665 | 0 | 0 |
| www.ck12.org | 43,206 | 0 | 0 | 14402 | 14402 | 0 | 0 | 0 | 0 | 0 | 0 | 14402 |
| www.elsevier.com | 36,504 | 0 | 0 | 36504 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| www.curriki.org | 13,104 | 0 | 0 | 2282 | 2282 | 0 | 0 | 6258 | 0 | 0 | 0 | 2282 |
| ocw.mit.edu | 6,275 | 0 | 0 | 0 | 0 | 0 | 0 | 6275 | 0 | 0 | 0 | 0 |
| www.teachersnotebook.com | 550 | 0 | 0 | 275 | 0 | 0 | 0 | 0 | 0 | 275 | 0 | 0 |
| www.tlsbooks.com | 7 | 0 | 0 | 5 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| repository.asu.edu | 7 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| epress.trincoll.edu | 5 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

**Table 7. Use of LRMI properties in educational domains in Corpus 2015**

| Domain | n-quads | educational Alignment | alignmentType | learning ResourceType | educational Use | educational Role | educational Framework | isBasedOnUrl | typical AgeRange | targetName | time Required | target Description | interactivity Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| brainpop.com | 484,139 | 83969 | 0 | 0 | 0 | 0 | 0 | 148263 | 83969 | 83969 | 0 | 83969 | 0 |
| phet.colorado.edu | 224,682 | 92146 | 92146 | 8078 | 24234 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8078 |
| www.merlot.org | 104,492 | 32584 | 32584 | 0 | 0 | 6740 | 0 | 0 | 0 | 32584 | 0 | 0 | 0 |
| www.bookshare.org | 66,171 | 0 | 0 | 8424 | 0 | 0 | 0 | 0 | 49323 | 0 | 0 | 0 | 8424 |
| www.enotes.com | 56,485 | 0 | 0 | 0 | 0 | 0 | 0 | 56485 | 0 | 0 | 0 | 0 | 0 |
| www.ck12.org | 38,994 | 0 | 0 | 12998 | 12998 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12998 |
| www.teacherspayteachers.com | 36,468 | 0 | 0 | 16547 | 0 | 0 | 0 | 0 | 16547 | 0 | 3374 | 0 | 0 |
| www.cteonline.org | 16,294 | 3287 | 3287 | 384 | 384 | 0 | 3287 | 0 | 0 | 3287 | 0 | 0 | 384 |
| www.whoi.edu | 8,722 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1994 | 8722 |
| law.stanford.edu | 6,098 | 0 | 0 | 6098 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| grad.arizona.edu | 4,912 | 0 | 4912 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ocw.mit.edu | 2,588 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2588 | 0 | 0 | 0 | 0 |
| www.turtlediary.com | 2,118 | 0 | 0 | 1059 | 0 | 0 | 0 | 0 | 1059 | 0 | 0 | 0 | 0 |
| www.oercommons.org | 1,982 | 0 | 148 | 863 | 0 | 224 | 0 | 0 | 409 | 169 | 0 | 169 | 0 |
| www.curriki.org | 1,420 | 0 | 0 | 355 | 355 | 0 | 0 | 0 | 355 | 0 | 0 | 0 | 355 |
| www.epubbooks.com | 1,356 | 0 | 0 |  | 0 | 0 | 0 | 0 | 0 | 0 | 1356 | 0 | 0 |
| www.getabstract.com | 806 | 0 | 0 | 403 | 0 | 0 | 0 | 0 | 0 | 0 | 403 | 0 | 0 |
| www.tlsbooks.com | 572 | 28 | 0 | 149 | 185 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 199 |
| www.audiobooks.com | 478 | 0 | 0 |  | 0 | 0 | 0 | 0 | 0 | 0 | 478 | 0 | 0 |
| www.teachersnotebook.com | 468 | 0 | 0 | 234 | 0 | 0 | 0 | 0 | 0 | 0 | 234 | 0 | 0 |
| 5ballov.qip.ru | 425 | 0 | 0 | 425 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| www.culture-formation.fr | 42 | 0 | 0 | 13 | 0 | 0 | 0 | 0 | 13 | 0 | 13 | 0 | 3 |
| www.owp.csus.edu | 37 | 0 | 0 | 10 | 9 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 9 |
| www.culture-formation.be | 36 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 12 | 0 | 12 | 0 | 0 |
| courses.p2pu.org | 32 | 0 | 0 | 28 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| education.lenardaudio.com | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 |
| epress.trincoll.edu | 20 | 4 | 0 | 4 | 4 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 4 |
| www.e-grammar.org | 12 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |

The property *targetUrl* was not detected in educational domains in any of the corpora, therefore it has been omitted from Table 6 and Table 7. The property *educationalFramework* has been omitted from Table 6 because it was not found in educational domains in Corpus 2014, as seen in Table 5. The total number of n-quads that include LRMI properties in educational domains of Corpus 2015 is

1,168,134, which represents 52.33% of the total n-quads for these properties (2,232,159, bottom of Table 5).

The total number of n-quads that include LRMI properties in educational domains in Corpus 2014 is 955,868, which represents 53.12% of the total n-quads for these properties (1,799,281, bottom of Table 5). The total number of n-quads that include LRMI properties in educational domains in Corpus 2015 is 1,168,134, which represents 52.33% of the total n-quads for these properties (2,232,159, bottom of Table 5). The share of educational domains in relation to the total number of domains that use these properties is almost the same in both corpora.

Regarding RQ2, the results of the use of LRMI properties in non-educational domains in Corpus 2014 and Corpus 2015 cannot be presented entirely in tables due to the high number of domains that use only a few properties. Therefore, we decided to present only the domains with the highest number of n-quads.

Table 8 presents a sample of the top non-educational domains and the properties they use in Corpus 2014. The purpose of each domain has been included next to its name. Only the properties that have n-quads in this sample are shown. The comma used to separate groups of thousands has been omitted in all columns of properties because of space constraints inside the table.

**Table 8**. Use of LRMI properties in non-educational domains in Corpus 2014

| Domain | n-quads | educational Alignment | alignmentType | learning ResourceType | isBasedOnUrl | typical AgeRange | time Required | interactivity Type |
|---|---|---|---|---|---|---|---|---|
| clarz.org (advertisement) | 469,608 | 0 | 0 | 0 | 469608 | 0 | 0 | 0 |
| www.fandango.com (entertainment) | 159,415 | 0 | 0 | 0 | 0 | 159415 | 0 | 0 |
| favim.com (image gallery) | 83,403 | 0 | 0 | 0 | 83403 | 0 | 0 | 0 |
| www.bcdb.com (cartoon database) | 67,097 | 0 | 0 | 0 | 0 | 0 | 67097 | 0 |
| www.sabah.com.tr (magazine from Turkey) | 15,285 | 0 | 0 | 0 | 0 | 5095 | 5095 | 5095 |
| www.bbc.co.uk (news and entertainment) | 1,794 | 846 | 846 | 0 | 0 | 0 | 0 | 0 |
| www.slideserve.com (web hosting) | 5,956 | 0 | 0 | 5956 | 0 | 0 | 0 | 0 |

Table 9 presents a sample of the top non-educational domains and the properties they use in Corpus 2015. The purpose of each domain has been included next to its name.

Moreover, to represent graphically the response to RQ2, Figure 2 presents a bar graph that shows a comparison of the use of LRMI properties in educational domains in both corpora. This bar graph shows the percentage of domains in which each property is used in relation to the total number of educational domains.

In Figure 2, the educational LRMI properties are shown in the top part of the bar graph and the general purpose LRMI properties are listed in the lower part of the bar graph, visually separated by a thick line in the table. The properties are presented in decreasing order of percentage.

Regarding the educational LRMI properties, the properties most used in educational domains in both corpora are *learningResourceType*, *educationalUse*, and *educationalAlignment*.

**Table 9**. Use of LRMI properties in non-educational domains in Corpus 2015

| Domain | n-quads | educational Alignment | alignmentType | learning ResourceType | educationalUse | isBasedOnUrl | typical AgeRange | time Required | interactivity Type |
|---|---|---|---|---|---|---|---|---|---|
| www.poemhunter.com (catalog of poems) | 539,235 | 0 | 0 | 0 | 0 | 0 | 179745 | 179745 | 179745 |
| www.staradvertiser.com (advertising) | 257,156 | 64289 | 64289 | 64289 | 0 | 0 | 0 | 0 | 64289 |
| www.penguin.com (entertainment) | 100,544 | 0 | | 0 | 0 | 0 | 100544 | 0 | 0 |
| www.fandango.com (entertainment) | 92,465 | 0 | 0 | 0 | 0 | 0 | 92465 | 0 | 0 |
| quesignifica.com (dictionary) | 19,048 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| www.bbc.co.uk (news and entertainment) | 15,944 | 0 | 0 | 0 | 0 | 0 | 0 | 15944 | 0 |
| drops.dagstuhl.de (online publication) | 4,770 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

This could be an expected result for the first two properties because these properties are meaningful when describing a resource. For instance, learningResourceType describes a type of educational resource with possible values such as "presentation," "lecture," "lesson plan," and "exercise." Similarly, educationalUse describes the recommended use for an educational resource in the pedagogical context such as "assignment," "group work," "test," and "self-learning."

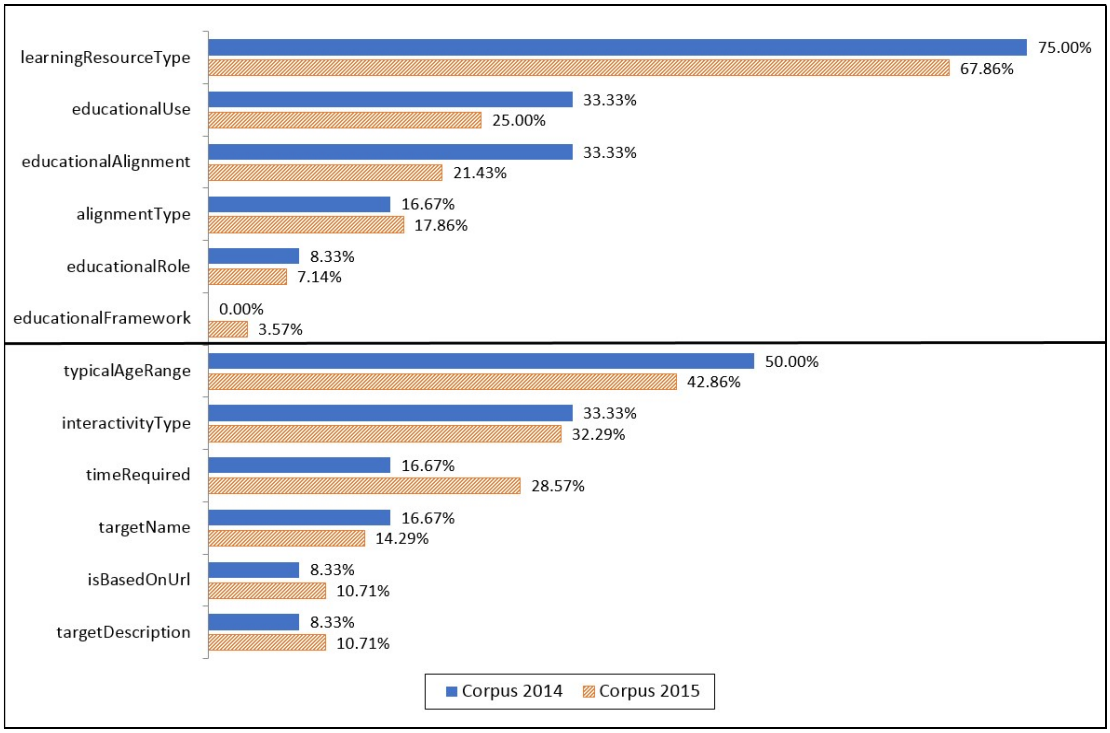


**Figure 2.** Percentage of domains that use each LRMI property in the total number of educational domains

In relation to RQ2, Figure 3 presents a comparison of the use of LRMI properties in non-educational domains in both corpora. This bar graph presents properties grouped in the same way as Figure 2. The graph shows the percentage of domains in which each property is used in relation to the total number of non-educational domains.

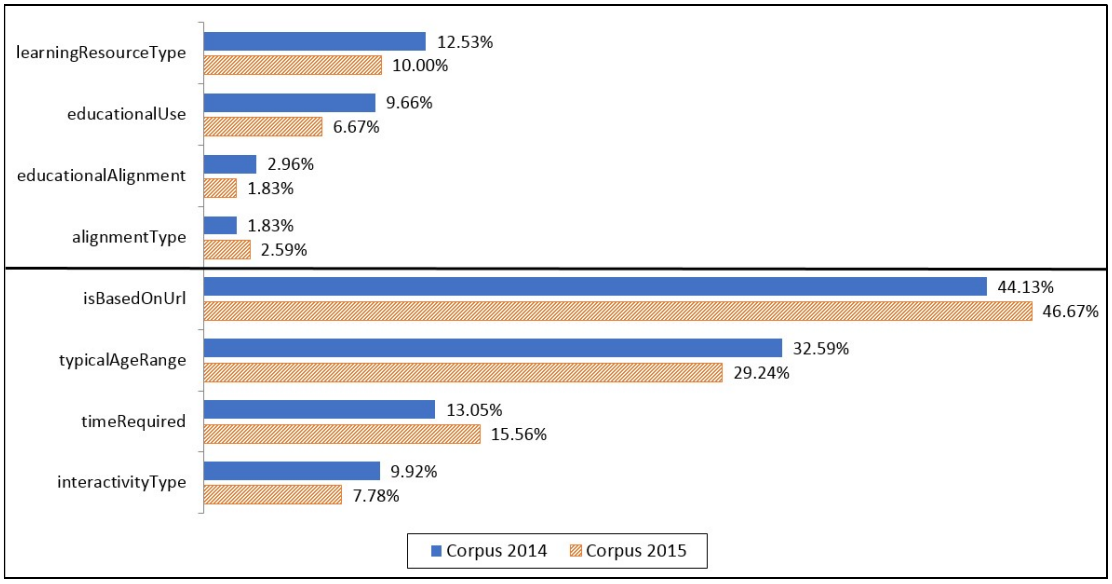


**Figure 3.** Percentage of domains that use each LRMI property in the total number of non-educational domains

The properties *educationalFramework*, *targetDescription*, *targetName*, and *educationalRole* are omitted from Figure 3 because there is no domain that uses them. The most used property is *isBasedOnUrl*, followed by *typicalAgeRange* and *timeRequired* because these properties can be used for many purposes.

Regarding RQ3, we found use of the educational LRMI properties aimed at educational content such as *learningResourceType*, *educationalUse*, *educationalAlignment*, and *alignmentType* in non-educational domains. This denotes that the appropriate use of these properties is not well known.

In non-educational domains in Corpus 2014, we found the following use of educational LRMI properties: e*ducationalAlignment* in eight domains; *educationalUse* in 18 domains; *learningResourceType* in 27 domains; *alignmentType* in seven domains; *educationalFramework* in one domain; and *educationalRole* in one domain. In non-educational domains in Corpus 2015, we found the following use of educational LRMI properties: e*ducationalAlignment* in seven domains; *educationalUse* in 37 domains; *learningResourceType* in 48 domains; *alignmentType* in seven domains; *educationalFramework* in one domain; and *educationalRole* in two domains.

Moreover, we detected a problem with the use of the *educationalAligment* property. This property is intended to describe the alignment with a defined educational framework, but we found incorrect values of this property, such as the strings "null" or "@en." We found that the property *learningResourceType* is misused to indicate the type of presentation of a printable book for sale (hardbound, paperback), while other properties present null values or unexpected values.

In addition, the properties *alignmentType* and *educationalFramework* should have been included together to meet the expected use defined by Schema.org [16], as seen in Table 3. Nevertheless, the number of domains that use *alignmentype* is higher than the number of domains that use *educationalFramework*, which is null in Corpus 2014 and a minimum value in Corpus 2015.

Furthermore, several important educational domains use only one property; for example, ocw.mit.edu only uses the property *typicalAgeRange* and law.standarford.edu only uses the property *learningResourceType*. This raises a problem because the properties used are insufficient for producing a helpful and adequate representation of educational resources.

Some properties such as *educationalAligment*, *educationalFramework*, *aligmentType*, *educationalRole*, *targetDescription*, *targetName*, and *targetUrl* are not used or are used minimally. The problems concerning the correct use of the educational LRMI properties are most likely derived from a lack of knowledge of their proper use in the educational field.

5.1.2 *Use of accessibility properties*

According to the analysis procedure, we obtained all domains that use accessibility properties in Corpus 2014 and Corpus 2015.

In relation to RQ1, in Corpus 2014, there are 1,151 n-quads that contain accessibility properties in 18 domains, while in Corpus 2015 there are 1,328,010 n-quads in 71 domains. These values show a huge growth in the use of accessibility properties for Microdata (more than a thousand times of n-quads and almost 300% in domains).

Regarding RQ1 and RQ2 in Corpus 2014, Table 10 shows a meaningful sample that includes the first six domains and the number of n-quads for each accessibility property. The rest of the domains have less than five n-quads. The purpose of each domain is included next to its name. Accessibility properties pending approval such as *accessMode*, *accessModeSufficient*, and *accessibilitySummary* were not detected at all, so they are not included in Table 10.

**Table 10.** Domains that use accessibility properties in Corpus 2014

| Domain | n-quads | accessibility API | accessibility Control | accessibility Feature | accessibility Hazard |
|---|---|---|---|---|---|
| youdescribe.org (audio description to YouTube videos) | 1,023 | 0 | 0 | 1,023 | 0 |
| www.totallypromotional.com (marketing) | 48 | 48 | 0 | 0 | 0 |
| www.pixellovegames.com (entertainment) | 41 | 0 | 41 | 0 | 0 |
| www.readybytes.net (software solutions) | 11 | 0 | 0 | 11 | 0 |
| www.rankya.com.au (SEO solutions) | 8 | 0 | 8 | 0 | 0 |
| tv.um.es (television of university campus) | 7 | 0 | 0 | 0 | 7 |

Furthermore, 12 out of 18 domains reported in Corpus 2014 use the *accessibilityFeature* property, which corresponds to 66.66%. This is an expected result because this is the most meaningful property to describe accessibility characteristics. The *accessibilityAPI* property is used by only one domain, most likely because technical knowledge is required to define the compatibility of the resource with the accessibility API.

In terms of RQ1 and RQ2 in Corpus 2015, Table 11 shows a meaningful sample that includes the first 12 domains and the number of n-quads for each accessibility property. The rest of the domains have less than 50 n-quads. The purpose of each domain is included next to its name. In Corpus 2015, two educational domains use accessibility properties. These domains are www.bookshare.org and openlibrary.org; both are included in Table 11. Accessibility properties pending approval such as *accessMode*, *accessModeSufficient*, and *accessibilitySummary* were not detected at all, so they are not included in Table 11.

**Table 11.** Domains that use accessibility properties in Corpus 2015

| Domain | n-quads | accessibility API | accessibility Control | accessibility Feature | accessibility Hazard |
|---|---|---|---|---|---|
| **www.bookshare.org (educational)** | 1,301,338 | 0 | 216,842 | 759,233 | 325,263 |
| forum.textpattern.com (open source CMS) | 21,012 | 3,502 | 7,004 | 0 | 10,506 |
| **openlibrary.org (educational)** | 1,815 | 0 | 0 | 1,815 | 0 |
| ottawamagazine.com (online magazine) | 1,112 | 0 | 0 | 1,112 | 0 |
| www.hunter-ed.com (sports) | 748 | 0 | 0 | 748 | 0 |
| www.telcodepot.com (phone services) | 618 | 0 | 67 | 507 | 44 |
| www.leitersburgcinemas.com (cinema) | 261 | 0 | 0 | 261 | 0 |
| www.readybytes.net (software solutions) | 182 | 0 | 0 | 182 | 0 |
| www.jonsatrom.com (artistic designer) | 222 | 0 | 0 | 0 | 222 |
| www.escolalivrededireito.com.br (business training) | 70 | 0 | 0 | 49 | 21 |
| www.rankya.com (SEO solutions) | 53 | 5 | 18 | 12 | 18 |
| blogspot.com (blogs) | 51 | 0 | 2 | 49 | 0 |

In Corpus 2015, there are 71 domains that use accessibility properties, of which 55 domains use the *accessibilityFeature* property, which corresponds to 77.46%. The lowest use is for the *accessibilityAPI* property, used by only eight domains (11.26%).

In response to RQ3, the prevailing value for *accessibilityFeature* in Corpus 2014 and Corpus 2015 is "audioDescription@en," i.e., the availability of auditory description in the English language for content, such as a video.

The prevailing value associated with *accessibilityAPI* is "ARIA@en" for all n-quads where it has been used. This value indicates the compatibility of the content with the accessibility standard Accessible Rich Internet Applications (WAI-ARIA). This standard allows the incorporation of additional semantics to the elements of the web interface to enhance accessibility through assistive technologies such as screen readers [56].

In terms of RQ4, when we examined the results of educational domains that use LRMI properties (Tables 6 and 7) and the use of accessibility properties (Tables 10 and 11), we found only one educational domain that uses both types of properties. This domain is **www.bookshare.org**, which belongs to an online library that offers books in digital format accessible to people with disabilities.

Some known educational domains, such as oercommons.org and merlot.org, include accessibility characteristics in their metadata to describe stored resources [57, 58]. However, we found that they use

Microdata with LRMI properties (as shown in Table 7), but they do not use Microdata with accessibility properties. This could be interpreted as a lack of awareness about the existence of accessibility properties for Microdata.

5.1.3 *Use of license properties*

The properties used to describe resource licenses are *license* from Schema.org and *useRightsUrl* from the LRMI specification (not adopted by Schema.org). The procedure analyzed the use of any of these properties and the type of license.

We examined the Creative Commons license, noted as "CC," and other types of license, which are noted as "Other." For example, the CC license is used in the domain www.merlot.org with the property *useRightsURL* and the value "http://creativecommons.org/licenses/by-nc-sa/3.0/us/" (implies a CC license under the following terms: Attribution-NonCommercial-ShareAlike 3.0 United States). The CC license is used in the domain www.ted.com with the property *license* and the value "http://creativecommons.org/licenses/by-nc-nd/3.0/" (implies a CC license under the following terms: Attribution-NonCommercial- NoDerivatives version 3.0). An "Other" license is used in the domain phet.colorado.edu with the property *useRightsURL* and the value "https://phet.colorado.edu/en/licensing/java" (Java licensing described for the University of Colorado).

Regarding RQ1 and RQ2, Table 12 presents the results for each property and type of license. The table includes the number of n-quads, the total number of domains that include each property and type of license, and the number of educational domains within them. The same information is presented for both corpora.

**Table 12.** Domains that use license properties in Corpus 2014 and Corpus 2015

| Property | Corpus 2014 | | | Corpus 2015 | | |
|---|---|---|---|---|---|---|
| | n-quads | Total domains | Educational domains | n-quads | Total domains | Educational domains |
| license (with CC) | 61,948 | 65 | 2 | 67,405 | 191 | 3 |
| license (with Other) | 2,550 | 27 | 0 | 533,946 | 59 | 1 |
| *Total use of the license property* | 64,498 | 92 | 2 | 601,351 | 250 | 4 |
| useRightsUrl (with CC) | 16,747 | 4 | 2 | 32,493 | 7 | 4 |
| useRightsUrl (with Other) | 14,187 | 2 | 1 | 29 | 2 | 0 |
| *Total use of the useRightsUrl property* | 30,934 | 6 | 3 | 32,522 | 9 | 4 |
| Total in Corpus (license + useRightsUrl) | 95,432 | 98 | 5 | 633,873 | 259 | 8 |

We found a small number of educational domains that use *license* or *useRightsUrl* properties in both corpora (five domains in Corpus 2014 and eight domains in Corpus 2015).

Concerning RQ2, when considering all the domains, the use of license property increases from 92 domains in Corpus 2014 to 250 domains in Corpus 2015, an increase of almost 172%. This is an expected result because the *license* property is part of Schema.org, while Schema.org did not adopt the *useRightsUrl* property. Thus, search engines such as Google do not recognize this property.

Further, we noticed a prevalence of the CC license in both corpora. This is an expected result because of the widespread use of this license on the web [59].

Tables 13 and 14 present a sample of educational and non-educational domains with the greater number of n-quads that use *license* or *useRightsUrl* properties, as well as the type of license in use (CC or Other).

Regarding RQ3, we found educational domains (ocw.mit.edu, phet.colorado.edu, www.merlot.org, and www.oercommons.org) in Corpus 2014 and Corpus 2015 that use the property *useRightsUrl*, which is not accepted by Schema.org. Hence, the Microdata annotation with this property is not recognized by the major search engines. This becomes a problem because of the high number of educational resources published by these domains.

**Table 13.** Domains that use license properties in Corpus 2014

| Domain | n-quads | Property | Type of license |
|---|---|---|---|
| **Educational domains** | | | |
| www.ted.com | 58,790 | license | CC |
| ocw.mit.edu | 14,609 | useRightsUrl | CC |
| phet.colorado.edu | 14,186 | useRightsUrl | Other |
| www.merlot.org | 2,126 | useRightsUrl | CC |
| www.curriki.org | 11 | license | CC |
| **Non-educational domains** | | | |
| www.lyricsbox.com | 1,762 | license | Other |
| blogspot.com | 1,661 | license | CC |
| dl.pconline.com.cn | 237 | license | CC |

**Table 14.** Domains that use license related properties in Corpus 2015

| Domain | n-quads | Property | Type of license |
|---|---|---|---|
| **Educational domains** | | | |
| www.bookshare.org | 147,000 | license | Other |
| www.ted.com | 59,339 | license | CC |
| phet.colorado.edu | 23,986 | useRightsUrl | CC |
| ocw.mit.edu | 6,425 | useRightsUrl | CC |
| www.merlot.org | 1664 | useRightsUrl | CC |
| oercommons.org | 416 | useRightsUrl | CC |
| era.library.ualberta.ca | 356 | license | CC |
| www.curriki.org | 355 | license | CC |
| **Non-educational domains** | | | |
| www.pond5.com | 382,324 | license | Other |
| blogspot.com | 3,842 | license | CC |
| www.referensimakalah.com | 761 | license | CC |

In terms of RQ4, we found several educational domains that use LRMI properties (Tables 6 and 7) and a license property (Tables 13 and 14): www.bookshare.org; phet.colorado.edu; ocw.mit.edu;

www.merlot.org; and www.curriki.org. However, considering the three sets of properties, only the domain www.bookshare.org uses LRMI properties, accessibility properties, and a license property.

### 5.2 *Errors detected in n-quads format*

The number of n-quads with format errors in Corpus 2014 is 36,690, which represents 0.0004% of the total n-quads (9,451,742,113) of this corpus. Similarly, the number of n-quads with format errors in Corpus 2015 is 16,048, which represents 0.0001% of the total n-quads (13,514,697,971) of this corpus. These n-quads were not considered in the analysis because of their minimal impact on the results of this research.

## 6 Conclusion

The evolution towards the semantic web has propelled the development of many technologies and standards. One of these is the embedded markup used to add semantic annotations to web content in conjunction with the controlled vocabulary of Schema.org, a standard driven by the most important search engines such as Google and Bing. However, the existence of a standard does not automatically translate into its application by professionals and practitioners. Therefore, it is important to appraise how these standards are actually used. In this research, we conducted a quantitative analysis of the extent of the adoption of Microdata for semantic annotations on educational resources and the results highlight their lack of use.

We based our analysis on two large datasets extracted from the Common Crawl Corpus by the WDC project. These corpora correspond to the crawling of December 2014 and November 2015. The size of the corpora for analysis enabled a reliable evaluation. Specifically, we evaluated the use of Microdata properties related to educational resources, considering the properties used to describe educational resources from the LRMI specification adopted by Schema.org, as well as the properties used to describe accessibility characteristics and resource licenses.

As explained in this paper, an indicator of the extent of the adoption of Microdata properties is the number of domains that use such properties. For this reason, the main findings from our quantitative analysis are presented in relation to RQ2: “How extensive is the use of each property belonging to the sets of LRMI properties, accessibility properties, and license properties in relation to the domains that use them? The results should consider educational and non-educational domains”:

- Concerning the LRMI properties, the number of domains that use such properties increased by 76.09%, from 46 domains in Corpus 2014 to 81 in Corpus 2015. As a part of these domains, the number of educational domains also increased by 115.4%, from 13 to 28 domains. These results show a growth in the adoption of Microdata in education.
- Concerning the use of accessibility properties, the number of domains that use such properties increased by 294.44%, from 18 domains in Corpus 2014 to 71 in Corpus 2015. Nevertheless, there were no educational domains in Corpus 2014 and only two educational domains were found in Corpus 2015.
- In terms of the properties related to the license of the resources, the number of domains that use such properties increased by 164.29%, from 98 domains in Corpus 2014 to 259 in Corpus

2015. The number of educational domains increased from five in Corpus 2014 to eight in Corpus 2015.

The results show that the number of domains that use these properties for Microdata is increasing. However, their use in educational resources is incipient.

Our qualitative analysis was guided by RQ3: "Have properties of the sets of LRMI properties, accessibility properties, and license properties been used for their expected purpose?":

- We found several problems, including the fact that educational LRMI properties are used by a minimal number of domains or are not used at all—for example, *educationalFramework* and *educationalRole* (see Figure 2). There is also a misuse of properties that should be used in conjunction with each other, such as *educationalFramework* and *alignmentType*. Further, educational LRMI properties specifically aim to describe the educational value of resources used in non-educational domains (see Figure 3), while we also found instances of wrong values of properties or properties with null or empty values. We also found educational domains that use only one educational LRMI property (*learningResourceType*) to describe educational resources, which is insufficient to properly describe the resource.
- On the other hand, when Schema.org adopted most of the LRMI properties, *useRightsUrl* was an exception. This means that this property is not a part of Schema.org; however, we found several important educational domains that use this property. This raises a problem because Microdata are intended to be processed by the general-purpose search engines that do not recognize this property.

To answer RQ4: "Have properties of the sets of LRMI properties, accessibility properties, and license properties been used together to improve the description of educational resources?":

- We found only one educational domain (www.bookshare.org) in Corpus 2015 that uses Microdata with LRMI properties to describe educational resources and includes accessibility characteristics and the license of resources all together. This domain belongs to an online library that offers books in digital format accessible to people with visual, physical, or learning disabilities (e.g., text to speech, increased sources, Braille printing). Therefore, it is expected that this domain would use accessibility properties.
- Considering the size of the corpora analyzed, this result regarding the use of Microdata annotations to improve educational resource searches is certainly discouraging.

Despite the results obtained, it is important to note that the use of Microdata can mean a breakthrough in relation to the problem of searching for educational resources on the web. Although at present, knowledge about the advantages provided by Microdata is not widespread, given the rise of HTML5 it is expected that Microdata adoption will continue to increase in all areas, including education.

Search engines currently produce personalized search applications (custom search) in different areas to deliver more meaningful search results. For instance, Google provides rich snippets to take advantage of Microdata and provides enriched information in areas such as music, personal information, and business, but this has still not been extended to educational resources or accessible web content.

A massive adoption of Microdata in educational domains may encourage Google and other search engines to include these areas in their custom search applications. To achieve this goal, it is important to improve knowledge about the advantages and the simplicity of Microdata adoption, as well as the proper use of the properties. Authoring tools could also contribute to this goal by providing an assisted process for adding Microdata annotations to resources' web pages.

For future work, we plan to complement this quantitative analysis with qualitative research that focuses on the most representative educational domains to identify the problems related to the adoption of Microdata and the comprehension of the Schema.org vocabulary.

## References


1. Sikos, L. F. Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data. Apress, 2015. DOI: 10.1007/978-1-4842-1049-9.
2. Meusel, R., Petrovski, P. and Bizer, C. The webdatacommons Microdata, RDFa and microformat dataset series. In: Proceedings of the 13th International Semantic Web Conference (ISWC '14) - Part I, Riva del Garda - Trentino, Italy, October 19-23, 2014, pp. 277-292. DOI: 10.1007/978-3-319-11964-9_18.
3. Guha, R., Brickley, R. and Macbeth, S. Schema.org: Evolution of Structured Data on the Web. Communications of the ACM, 2016; 59(2): 44-51. DOI: 10.1145/2857274.2857276.
4. Piedra, N., Chicaiza, J., López-Vargas, J. and Caro, E.T. Seeking Open Educational Resources to Compose Massive Open Online Courses in Engineering Education: An Approach based on Linked Open Data. Journal of Universal Computer Science, 2015; 21(5): 679-711. DOI: 10.3217/jucs-021-05-0679.
5. Navarrete, R. and Luján-Mora, S. (2014) Metadata in Open Educational Resources websites: a review from the perspective of disabled users' requirements. In: Proceedings of the 6th International Conference on Education and New Learning Technologies (EDULEARN), Barcelona, Spain, July 7- 9, 2014, pp. 111-120. Available at: https://goo.gl/2MV6TA.
6. Allen, E. and Seaman, J. Opening the Curriculum: Open Educational Resources in U.S. Higher Education. Babson Survey Research Group, 2014, pp. 29-30.
7. Yu, L. A Developer's Guide to the Semantic Web. Springer Berlin Heidelberg, 2014. DOI: 10.1007/978-3-662-43796-4_10.
8. Navarrete, R. and Luján-Mora, S. Evaluating findability of Open Educational Resources from the perspective of users with disabilities: A preliminary approach. In: Proceedings of the Second International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, April 8-10, 2015, pp. 112-119. DOI: 10.1109/ICEDEG.2015.7114457.
9. UNESCO. World Education Forum, https://en.unesco.org/world-education-forum-2015/incheon-declaration (2015, accessed September 2017).
10. World Health Organization. World Report on Disability, https://goo.gl/q88CuW (2011, accessed September 2017).
11. United Nations. World Population Ageing, https://goo.gl/g3tU7U (2015, accessed September 2017).
12. Hawksey, M., Barker, P. and Campbell, L.M. New Approaches to Describing and Discovering Open Educational Resources. In: Proceedings of OER13: Creating a Virtuous Circle, Nottingham, England, March 26-27, 2013. Available at: http://publications.cetis.ac.uk/2013/767.
13. Haas, K., Mika, P., Tarjan, P. and Blanco, R. Enhanced Results for Web Search. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2011, Beijing, China, pp. 725-734. DOI: 10.1145/2009916.2010014.
14. Meusel, R., Bizer, C. and Paulheim, H. A Web-scale Study of the Adoption and Evolution of the schema.org Vocabulary over Time. In: Proceedings of the ACM 5th International Conference on Web Intelligence, Mining and Semantics (WIMS '15), Larnaca, Cyprus, July 13-15, 2015, pp. 1-11. DOI: 10.1145/2797115.2797124.
15. Schema.org. What is schema.org, https://schema.org/ (2015, accessed November 2016).

16. Learning Resource Metadata Initiative. LRMI Version 1.1, http://lrmi.dublincore.net/lrmi-1-1/ (2014, accessed November 2016).
17. Common Crawl Foundation. Common Crawl, http://commoncrawl.org/ (accessed November 2016).
18. Data and Web Science Research Group - University of Manheim. Web Data Commons, http://webdatacommons.org/ (2013, accessed November 2016).
19. Taibi, D. and Dietze, S. Towards Embedded Markup of Learning Resources on the Web: An Initial Quantitative Analysis of LRMI Terms Usage. In: Proceedings of the 25th International Conference Companion on World Wide Web, Montreal, Canada, April 11-15, 2016, pp. 513-517. DOI: 10.1145/2872518.2890464.
20. Sahoo, P., Gadiraju, U., Yu, R., Saha, S. and Dietze, S., Analysing Structured Scholarly Data embedded in Web Pages, Semantics, Analytics, Visualisation: Enhancing Scholarly Data (SAVE-SD2016), co-located with the 25th International World Wide Web Conference, Montreal, Canada, April 11, 2016. Available at: https://goo.gl/NdLCTe.
21. W3C. HTML5, http://www.w3.org/TR/html5/ (2015, accessed November 2016).
22. W3C. HTML Microdata, https://www.w3.org/TR/Microdata/ (2015, accessed November 2016).
23. Paulheim, H. What the Adoption of schema.org Tells About Linked Open Data. In: Proceedings of the 5th International Workshop on Using the Web in the Age of Data (USEWOD '15) and the 2nd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '15). 2015, Portoroz, Slovenia, June 1, 2015, pp. 85-90. Available at: http://ceur-ws.org/Vol-1362/PROFILES2015_paper6.pdf.
24. Bizer, C., Heath, T. and Berners-Lee, T. Linked data-the story so far. International Journal on Semantic Web and Information Systems, 2009; 5(3): 1-22. DOI: 10.4018/jswis.2009081901.
25. Fons, T., Penka, J. and Wallis, R. OCLC's Linked Data Initiative: Using Schema.org to Make Library Data Relevant on the Web. Information Standards Quarterly Spring/Summer, 2012; 2(3): 1-6. DOI: 10.3789/isqv24n2-3.2012.05.
26. Ronallo, J. HTML5 Microdata and Schema.org. Code4Lib, 2012; 16: 1-4. Available at: http://journal.code4lib.org/articles/6400.
27. Patel-Schneider, P. F. Analyzing Schema.org. In: Proceedings of the 13th International Semantic Web Conference (ISWC '14) - Part I, Riva del Garda - Trentino, Italy, October 19-23, 2014, pp. 261-276. DOI: 10.1007/978-3-319-11964-9_17.
28. Barker, P. and Campbell, L. Learning Resource Metadata Initiative: using schema.org to describe open educational resources. In: Proceedings of OpenCourseWare Consortium Global 2014: Open Education for a Multicultural World, Ljubljana, Slovenia, April 23 - 25, 2014, pp. 1-4. Available at: http://publications.cetis.org.uk/wp-content/uploads/2014/09/Paper_34-LMRI1.pdf.
29. Levy, Y. and Ellis, T. J. A systems approach to conduct an effective literature review in support of information systems research. Informing Science: International Journal of an Emerging Transdiscipline, 2006; 9(1): 181-212. Available at: https://goo.gl/4tCNPq.
30. Pastore, S. Website development and web standards in the ubiquitous world: where are we going? WSEAS Transactions on Computers, 2012; 11(9): 309-318. Available at: https://goo.gl/fDFZpa.
31. Pohorec, S., Zorman, M. and Kokol, P. Analysis of approaches to structured data on the web. Computer Standards & Interfaces, 2013; 36(1): 256-262. DOI: 10.14778/2180912.2180920.
32. Wu, Z., Xu, Y., Zhang, C., Yang, Y. and Ji, Y. (2016) Towards Semantic Web of Things: From Manual to Semi-automatic Semantic Annotation on Web of Things. In: Proceedings of the 2nd International Conference (BigCom), Shenyang, China, July 29-31, 2016, pp. 295-308. DOI: 10.1007/978-3-319-42553-5_25.
33. Hilliker, R. J., Wacker, M. and Nurnberger, A. L. Improving Discovery of and Access to Digital Repository Contents Using Semantic Web Standards: Columbia University's Academic Commons. Journal of Library Metadata, 2013; 13(2-3), 80-94. DOI: 10.7916/D86M34RR.
34. Flotyński, J. and Walczak, K. (2013, September). Microformat and Microdata schemas for interactive 3d web content. In: Proceedings of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS), Kraków, Poland, September 8-11, 2013, pp. 549-556. Available at: https://annals-csis.org/proceedings/2013/pliks/231.pdf.

35. Stoll, K. U., Ge, M. and Hepp, M. Understanding the Impact of E-Commerce Software on the Adoption of Structured Data on the Web. In: 16th International Conference on Business Information Systems, Poznań, Poland, June 19-20, 2013, pp. 100-112. DOI: 10.1007/978-3-642-38366-3_9.
36. Sikos, L. F. Advanced (X) HTML5 metadata and semantics for Web 3.0 videos. DESIDOC Journal of Library & Information Technology, 2011; 31(4): 247-252. DOI: 10.14429/djlit.31.4.1105.
37. Kutuzov, A. and Ionov, M. Untangling the Semantic Web: Microdata use in Russian video content delivery sites. In: International Conference on Analysis of Images, Social Networks and Texts (AIST), Yekaterinburg, Russia, April 10-12, 2014, pp. 274-279. DOI: 10.1007/978-3-319-12580.
38. Pabitha, P., Vignesh Nandha Kumar, K., Pandurangan, N., Vijayakumar. R. and Rajaram, M. Semantic Search in Wiki using HTML5 Microdata for Semantic Annotation. International Journal of Computer Science Issues, 2011; 8(3): 388-394. Available at: https://goo.gl/yVWbr1.
39. Lars, J. (2012). HTML5, MICRODATA AND SCHEMA.ORG - Towards an Educational Social-semantic Web for the Rest of Us? In: Proceedings of the 4th International Conference on Computer Supported Education (CSEDU), Volume 1, Porto, Portugal, April 16-18, 2012, pp. 101-104. DOI: 10.5220/0003895901010104.
40. Matosevic, G. The Adoption of Semantic Annotations of Products in Web Shops. International Journal of Computer and Communication Engineering, 2014; 3(1): 6-10. DOI: 10.7763/IJCCE.2014.V3.282.
41. Meusel R, Primpeli A, Meilicke C, Paulheim, H. and Bizer, C. Exploiting Microdata Annotations to Consistently Categorize Product Offers at Web Scale. In: Proceedings of the 16th International Conference on Electronic Commerce and Web Technologies (EC-Web), Valencia, Spain, September 3-4, 2015, pp. 83-99. DOI: 10.1007/978-3-319-27729-5_7.
42. Hepp, M. The Web of Data for E-Commerce: Schema.org and GoodRelations for Researchers and Practitioners. In: Proceedings of the 15th International Conference on Engineering the Web in the Big Data Era, Volume 9114, Rotterdam, The Netherlands, June 23 - 26, 2015, pp. 723-727. DOI: 10.1007/978-3-319-19890-3_66.
43. Ristoski P. and Mika P. Enriching Product Ads with Metadata from HTML Annotations. In: The Semantic Web. Latest Advances and New Domains, Volume 9678, Heraclion, Crete, Greece, May 29-June 1. 2016, pp.151-167. DOI: 10.1007/978-3-319-34129-3_10.
44. Nogales, A., Sicilia, M. A., Sánchez-Alonso, S. and Garcia-Barriocanal, E. Linking from Schema.org Microdata to the Web of Linked Data: An empirical assessment. Computer Standards & Interfaces, 2016, 45: 90-99. DOI: 10.1016/j.csi.2015.12.003.
45. DiFranzo, D., Erickson, J. S., Gloria, M. J. K. T., Luciano, J. S., McGuinness, D. L. and Hendler, J. The web observatory extension: facilitating web science collaboration through semantic markup. In: Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, April 7 -10, 2014, pp. 475-480. DOI: 10.1145/2567948.2576936.
46. Mika, P. and Potter, T. Metadata statistics for a large web corpus. In: Bizer C, Heath T, Berners-Lee T, et al. (eds) CEUR Workshop on Linked Data on the Web (LDOW), Lyon, FR, April 16, 2012, pp. 6-10. Available at: http://ceur-ws.org/Vol-937/ldow2012-inv-paper-1.pdf.
47. Mühleisen, H. and Bizer, C. Web data commons - Extracting structured data from two large web corpora. In: Bizer C, Heath T, Berners-Lee T, et al. (eds) CEUR Workshop on Linked Data on the Web (LDOW), Lyon, France, April 16, 2012, pp. 2-5. Available at: https://goo.gl/eDyi8V.
48. Bizer, C., Eckert, K., Meusel, R., Mühleisen, H., Schuhmacher, M. and Völker, J. Deployment of RDFa, Microdata, and Microformats on the Web - A Quantitative Analysis. In: Proceedings of the 12th International Semantic Web Conference (ISWC 2013) - Proceedings Part II, Sydney, Australia, October 21-25, 2013, pp. 17-32. DOI: 10.1007/978-3-642-41338-4_2.
49. Mühleisen, H. Vocabulary Usage by Pay-Level Domain, https://goo.gl/787pCn (2015, accessed September 2017).
50. Web Data Commons. Download Instructions for the WDC RDFa, Microdata, and Microformats Data Sets, https://goo.gl/9JBCtH (2014, accessed September 2017).
51. RDFa, Microdata, Embedded JSON-LD, and Microformats Data Sets - November 2015, https://goo.gl/rccIRV (2015, accessed October 2016).

52. W3C. RDF 1.1 N-Quads, https://www.w3.org/TR/n-quads/ (2014, accessed September 2017).
53. Schema.org. Hosted extension: pending, https://pending.schema.org/ (2015, accessed, September 2017).
54. DMOZ Internet Directory, http://dmozlive.com/ (accessed June 2017).
55. Schema org. CreativeWork, http://schema.org/CreativeWork (2015, accessed September 2017).
56. W3C. Accessible Rich Internet Applications (WAI-ARIA) 1.0, https://www.w3.org/TR/wai-aria (2014, accessed September 2017).
57. Navarrete, R. and Luján-Mora, S. Accessibility considerations in Learning Objects and Open Educational Resources. In: Proceedings of the 6th International Conference of Education, Research and Innovation (ICERI), Seville, Spain November 18-20, 2013, pp. 521-530. Available at: https://goo.gl/H82jxs.
58. Navarrete, R. and Luján-Mora, S. Evaluating accessibility of Open Educational Resource website with an heuristic method. In: Proceedings of the 9th International Technology, Education and Development Conference (ITHET), Caparica - Lisbon, Portugal June 11-13, 2015, pp. 6402-6412. Available at: https://goo.gl/fHE5at.
59. Creative Commons. State of the commons, https://stateof.creativecommons.org/2015/ (2015, accessed September 2017).