# Analysing and Improving Embedded Markup of Learning Resources on the Web

Stefan Dietze
L3S Research Center
30176 Hannover, Germany
dietze@l3s.de

Davide Taibi
CNR Institute for Educational Technologies, Palermo, Italy
davide.taibi@itd.cnr.it

Ran Yu
L3S Research Center
30176 Hannover, Germany
yu@l3s.de

Phil Barker
Heriot Watt University
Edinburgh, United Kingdom
phil.barker@hw.ac.uk

Mathieu d'Aquin
The Open University
Milton Keynes, United Kingdom
m.daquin@open.ac.uk

## ABSTRACT

Web-scale reuse and interoperability of learning resources have been major concerns for the technology-enhanced learning community. While work in this area traditionally focused on learning resource metadata, provided through learning resource repositories, the recent emergence of structured entity markup on the Web through standards such as RDFa and Microdata and initiatives such as schema.org, has provided new forms of entity-centric knowledge, which is so far under-investigated and hardly exploited. The Learning Resource Metadata Initiative (LRMI) provides a vocabulary for annotating learning resources through schema.org terms. Although recent studies have shown markup adoption by approximately 30% of all Web pages, understanding of the scope, distribution and quality of learning resources markup is limited. We provide the first public corpus of LRMI extracted from a representative Web crawl together with an analysis of LRMI adoption on the Web, with the goal to inform data consumers as well as future vocabulary refinements through a thorough understanding of the use as well as misuse of LRMI vocabulary terms. While errors and schema misuse are frequent, we also discuss a set of simple heuristics which significantly improve the accuracy of markup, a prerequisite for reusing learning resource metadata sourced from markup.

## Keywords

LRMI, schema.org, Web markup, learning resources

## 1. INTRODUCTION

Exploitation of learning resources on the Web has been a major concern for the technology enhanced learning community in both research and practice. Traditionally, work in this area has focused on providing repositories of learning resource metadata, where supporting technologies are concerned with increasing interoperability through metadata standards [9], linking of resources and vocabularies through Linked Data techniques [8] or retrieval and recommendation techniques. However, while a significant amount of resources has been exposed and annotated on the Web, reuse is still lacking. This can be attributed to shortcomings prevalent in many data sharing efforts on the Web, such as lack of quality of resources as well as their annotations, diversity of metadata standards and vocabularies, accessibility of data [11] and the often poorly maintained metadata descriptions, raising concerns with respect to trust and reliability when attempting to reuse third-party data and resources.

More recently, entity-centric annotations embedded in HTML pages have become prevalent on the Web, building on standards such as RDFa[1], Microdata[2] and Microformats[3]. In this context, the *Learning Resource Metadata Initiative (LRMI)*[4] provides a vocabulary to enable markup of (online) learning resources through *schema.org* terms. Driven by the support from search engine providers such as Google, Yahoo!, Bing and Yandex, schema.org markup has reached significant adoption, where more than 30% of all Web pages already provide some form of markup [1]. As such, markup constitutes a source of entity-centric data on the Web at an unprecedented scale.

Even though the general adoption of schema.org respectively markup suggests the significant availability of learning-related markup data on the Web, understanding of the scope, distribution and quality of learning resource markup and related entities is limited so far. In particular, while schema.org provides recommendations of terms and their use, it does not represent a formal and deeply constrained ontological framework [4], and hence use and interpretation of terms varies heavily. For instance, while the schema.org property *author*[5] expects a value range of instances of type *Person* or *Organisation*, most commonly literals are used instead [2]. Given the widespread use of vocabulary terms in unintended ways, understanding the use and repurposing of vocabulary terms is crucial to (i) enable the reuse of markup

WWW 2017 Companion, April 3-7, 2017, Perth, Australia. ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054160

[1] https://www.w3.org/TR/rdfa-syntax/
[2] https://www.w3.org/TR/microdata/
[3] http://microformats.org/
[4] http://lrmi.dublincore.net/
[5] https://schema.org/author

data, for instance, as part of learning resource recommenders and search engines or to build targeted knowledge graphs, and (ii) to inform the shape of further vocabulary extensions and refinements. Understanding successfully adopted terms and modelling patterns can guide future vocabulary improvements within both the schema.org and LRMI communities.

In this work, we provide the first large-scale analysis of LRMI adoption on the Web, investigating the adoption, evolution, distribution and scope of LRMI markup and co-occurring entity annotations. To ensure a representative study, we use the largest publicly available Web crawl, i.e. LRMI markup extracted from the Common Crawl[6] from three consecutive years (2013-2015), with approximately 2 billion crawled Web pages per year. While errors and schema misuse are increasingly frequent, we also introduce a set of cleansing techniques which significantly improve the accuracy of markup, a prerequisite for reusing learning resource metadata sourced from markup. As part of this work, we address the following research questions: *RQ1)* How did the adoption of LRMI terms evolve over time and what are successfully adopted terms and associated modelling pattern?, *RQ2)* How is learning resource metadata, as expressed through LRMI markup, distributed across the Web and how did such distribution change over time?, and *RQ3)* What is the quality of LRMI markup and how can frequent errors be improved so that data can easily be interpreted and reused?

By addressing the aforementioned research questions, we inform data consumers, the future vocabulary design process as well as data providers through the following main contributions:

*(i)* a large-scale dataset of LRMI markup extracted from representative Web crawls of three consecutive years,
*(ii)* a first large-scale study of LRMI adoption and quality on the Web,
*(iii)* a set of techniques for cleansing and improving LRMI data, in order to aid data reuse and interpretation.

The following section introduces background and related work, while Section 3 describes the methodology and dataset. Sections 4 and 5 present the analysis and results addressing RQ1 and RQ2 respectively, while RQ3 is addressed in Section 6 through an assessment of LRMI markup quality and the proposal of a set of heuristics and data cleansing techniques. Section 7 finally discusses the findings and lessons learned, while Section 8 concludes the paper and discusses potential future work.

## 2. BACKGROUND & RELATED WORK

Sharing of learning resources on the Web through some form of metadata has been subject of extensive research and practice throughout the past decade, leading to a variety of metadata standards, vocabularies and schemas for annotating learning resources and designs of varying granularity [9], such as *IEEE LOM*[7], *IMS Learning Resource Metadata*[8] or *ADL SCORM*[9]. While the lack of reuse and cross-standard interoperability as well as the limited use of shared vocabularies prevented Web-scale interoperability and take-up across distinct repositories or platforms [12], more recent efforts have adopted RDF and Linked Data-based approaches to improve the linking, understanding and integration of learning resources annotations [8], culminating into platforms such as the *LinkedUp Data Catalog*[10] and a wealth of Linked Data-compliant data collections [5][8]. However, cross-domain issues inherent to (linked) data sharing initiatives [11] in general as well as learning-related linked data in particular include the highly heterogeneous quality of data and resources, the lack of currentness, dynamics, availability and accessibility [3] as well as scalability and performance issues, leading to a limited uptake and reuse of available data.

On the other hand, embedded Web page markup vocabularies such as schema.org emerged as a means to embed entity-centric data directly into Web pages to be used by major search engines to interpret Web content. The Web Data Commons [1], a recent initiative investigating the Common Crawl, i.e. a Web crawl of approximately 2 billion HTML pages from over 15 million pay-level-domains (PLDs) found that 30% of all pages contain some form of embedded markup already, resulting in a corpus of 20.48 billion RDF quads in 2014. Considering the upward trend of adoption - the proportion of pages containing markup increased from 5.76% to 30% between 2010 and 2014 - and the still comparably limited nature of the investigated Web crawl, the scale of the data suggests potential for a range of tasks, such as entity retrieval, knowledge base population, or entity summarization [7].

With respect to educational resources, the schema.org extension developed by the *Learning Resource Metadata Initiative (LRMI)* has been included into schema.org in April 2013 and is currently under development by the LRMI task force of the *Dublin Core Metadata Initiative (DCMI)*[11]. In particular, the following LRMI predicates for the description of educational characteristics of creative works (*s:CreativeWork*) are part of schema.org and investigated here: *educationalAlignment*, *educationalUse*, *timeRequired*, *typicalAgeRange*, *interactivityType*, *learningResourceType*, *isBasedOnUrl*. In addition, two LRMI-specific types are part of the schema.org vocabulary: *AlignmentObject* and *EducationalAudience*. With 'terms' we refer to both properties and types in the following. While an early study [13] provided initial insights into the significant adoption of LRMI, this work was based on a limited and outdated dataset, considering a subset of the Common Crawl only. Also, there had been no attempts to deal with the inherent data quality problem when dealing with markup data. These shortcomings are elevated by the fact that data extracted from markup has fundamentally distinct characteristics compared to traditional Linked Data, consisting of vast amounts of flat, disconnected and often redundant entity descriptions [6].

## 3. DATASET & METHODOLOGY

We exploit the structured data corpus of the Web Data Commons, containing all Microformat, Microdata and RDFa data from the Common Crawl (*CC*). In particular, as the LRMI metadata schema has been released since 2013, we have considered the data extracted from the releases of November 2013 ($CC_{13}$), December 2014 ($CC_{14}$) and November 2015 ($CC_{15}$) of the Common Crawl. In particular, we refer to the following datasets:

- $CC=\{CC_{13}, CC_{14}, CC_{15}\}$ refers to the Common Crawl, where, for instance, $\mathrm{CC}_{14}$ refers to the set of all documents *d* contained within the December 2014 release of the Common Crawl.

[6] http://commoncrawl.org/

[7] http://grouper.ieee.org/groups/ltsc/wg12/20020612-Final-LOM-Draft.html

[8] https://www.imsglobal.org/metadata/index.html

[9] http://www.adlnet.org

[10] http://data.linkededucation.org/linkedup/catalog/

[11] http://wiki.dublincore.org/index.php/AB-Comm/ed/LRMI/TG

- $M=\{M(CC_{13}), M(CC_{14}), M(CC_{15})\}$ refers to the markup extracted from the respective Common Crawls, where $M(CC_{14})$ refers to the markup extracted from the 2014 Common Crawl release introduced above and is provided through the structured data corpus of the respective Web Data Commons[12] release. In $M$, each entity description corresponds to a set of quadruples $q$ of the form $\{s, p, o, u\}$, where $s, p, o$ represent a triple consisting of subject, predicate, object and $u$ represents the URL of the document $d$ from which the triple has been extracted respectively. For a particular real-world entity $e$, usually there exist $n \geq 0$ subjects $s$ which represent distinct descriptions of $e$.
- $LRMI=\{LRMI(CC_{13}), LRMI(CC_{14}), LRMI(CC_{15})\}$ refers to the LRMI markup extracted from the Common Crawl (*CC*), respectively 2013, 2014, 2015 releases of *CC*. Precisely, this dataset contains all embedded markup statements extracted from documents (in the respective CC corpus) which contain at least one triple $\{s, p, o\}$ where either $p$ refers to any of the LRMI predicates or $s$ or $o$ represent instances of LRMI-specific types *AlignmentObject* or *EducationalAudience* described in Section 2. While LRMI markup is a specific set of Web markup, the LRMI corpus of a respective year is a subset of the corresponding markup corpus, e.g. $LRMI(CC_{14}) \subseteq M(CC_{14})$.
- $LRMI'=\{LRMI'(CC_{13}), LRMI'(CC_{14}), LRMI'(CC_{15})\}$ refers to a variant of the LRMI corpus denoted above, where additionally quads were included which contained erroneous LRMI statements, considering the frequent errors described in [10], for instance, quads involving misspellings of LRMI terms (see Section 6).

Table 1 provides an overview of the size of investigated datasets, namely the amount of documents ($|D|$), URLs ($|U|$) and quads ($|Q|$). Note that the values in brackets indicate the relative proportion of URLs in $CC_i$ which provide markup ($M(CC_i)$) respectively LRMI markup ($LRMI(CC_i)$). While the *M* and *CC* corpora are available through the Web Data Commons and the Common Crawl websites, we made available the *LRMI* and *LRMI'* datasets together with other resources[13]. Note that Section 4 and 5 investigate *LRMI* corpora only, in order to provide an accurate analysis of LRMI adoption, while Section 6 investigates frequent errors within *LRMI'*.

**Table 1. Sizes of datasets under investigation**

| | 2013 | 2014 | 2015 |
|---|---|---|---|
| $\|D\|$ where $d \in CC_i\|$ | 2,224,829,946 | 2,014,175,679 | 1,770,525,212 |
| $\|U\|$ where $u \in M(CC_i)\|$ | 585,792,337 (26.3%) | 620,151,400 (30.7%) | 541,514,775 (30.5%) |
| $\|U\|$ where $u \in LRMI(CC_i)$ | 83,791 (0.00003766%) | 430,861 (0.00021391%) | 779,260 (0.00044012%) |
| $\|U\|$ where $u \in LRMI'(CC_i)$ | 84,098 | 430,895 | 929,573 |
| $\|Q\|$ where $q \in M(CC_i)$ | 17,241,313,916 | 20,484,755,485 | 24,377,132,352 |
| $\|Q\|$ where $q \in LRMI(CC_i)$ | 9,245,793 (0.00053625%) | 26,256,833 (0.00128177%) | 44,108,511 (0.00180942%) |
| $\|Q\|$ where $q \in LRMI'(CC_i)$ | 9,251,553 (0.00053659%) | 26,258,524 (0.00128185%) | 69,932,849 (0.00286878%) |

## 4. ADOPTION OF LRMI TERMS

In order to address *RQ1*, we conducted an analysis of LRMI term adoption and its evolution. Figure 1 depicts the occurrence of LRMI vocabulary terms in $LRMI(CC_i)$, where occurrence refers to the number of statements involving any of the LRMI terms under investigation (Section 2). A further analysis of the growth rate of the URLs and statements containing particular LRMI terms in $LRMI(CC_i)$ compared to the previous year, i.e. $LRMI(CC_{i-1})$ is provided online[13], in order to provide a better understanding of the evolution of individual terms.

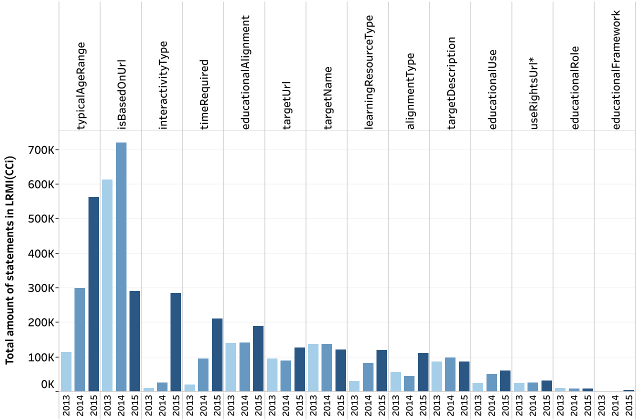

**Figure 1. Total number of statements in LRMI(CCi) involving particular LRMI terms in 2013-2015**

As shown, particular properties such as *typicalAgeRange* and *interactivityType* have reached comparably wide adoption in 2015 and show significant growth, while others, for instance, *targetName*, *targetUrl* and *targetDescription*, show fairly static adoption for all three years. One explanation here is the general-purpose nature of the former attributes, which are potentially applied to all sorts of informal learning resources, while the latter properties directly aim at aspects related to formal educational material and corresponding educational frameworks. As shown in Section 5, the majority of annotations refer to informal learning resources, such as videos or tutorials, where aspects of formal education do not apply. This is also underlined by the limited occurrence of *educationalFramework*, which has been absent entirely in 2013 and 2014. Another observation concerns the drop in use of *isBasedOnUrl*, caused by its deprecation in 2014, demonstrating the adoption of vocabulary evolution.

**Table 2. Datatype property usage**

| | # quads (transversal) | % datatype | % literals |
|---|---|---|---|
| 2013 | 7,251,417 | 55,82 | 69,52 |
| 2014 | 19,916,701 | 56,06 | 78,07 |
| 2015 | 46,883,557 | 64,39 | 96,82 |

Inline with the observation that simple datatype properties appear to see wider acceptance, Table 2 shows the total amount of transversal quads (in *LRMI'*), i.e. quads involving non-hierarchical properties such as *rdf:type*, the proportion (%) of datatype properties and the proportion of statements which actually refer to literals, i.e. are used as simple datatype statements. The figures underline the widespread use of literal-based statements (>96% in 2015), even when using object properties. This underlines a lack of acceptance of controlled vocabularies or more complex modelling patterns and leads to large amounts of flat resource and entity descriptions as opposed to an interconnected graph structure.

On the other hand, the use of *educationalAlignment* and *alignmentType* has seen a significant growth in 2015, what appears to indicate a limited yet increasing adoption of the

[12] http://webdatacommons.org

[13] Datasets and further resources are available from http://lrmi.itd.cnr.it

modelling concept behind LRMI, where any resource (*CreativeWork*) can be associated with learning-related properties. While this concept is less straightforward from a markup provider's perspective, it reflects the general understanding of learning resources as arbitrary knowledge resources which may or may not be used in a learning context. Additional observations include the largest growth for the property *interactivityType* and a growth of more than 100% for 8 out of 12 terms.

Generalising about the characteristics of terms which have seen wide adoption, it becomes apparent that particular modeling pattern seem to be more successful than others. Specifically, terms which are either highly general, such as *typicalAgeRange*, and/or are simple data-type properties which expect literal values (such as *name*) seem to be among the most frequently used. A pattern which mirrors the general findings of [1] in the LRMI context and yields implications for both data consumers as well as future directions for the extension of LRMI as discussed in Section 7. Additional data is presented on our resources Website[13], including high resolution figures and detailed statistics of the adoption of terms as part of entities, quads and documents.

## 5. DISTRIBUTION OF LRMI DATA

This section investigates the distribution of LRMI markup across providers (pay-level-domains, PLDs), and top-level-domains (TLDs). Table 3 shows the total number of PLDs within $CC_i$ and *LRMI*($CC_i$), indicating a significant increase of markup providers in general ($CC_i$) as well as for LRMI markup. While some particular providers, e.g. *lap.hu*, provide a significant amount of independently maintained subdomains which, however, do not constitute PLDs, these were excluded. The total amount of domains (PLDs and subdomains) is 3,659 in 2015.

As shown in Figure 2, which depicts the amount of documents (log scale) per PLD, LRMI data is spread across PLDs following a power-law distribution, with the top 10% of providers contributing 98.4% of all documents containing markup statements. This correlates directly with the amount of pages and resources of a particular markup-providing Website/PLD.

**Table 3. Total number of PLDs in $CC_i$, M($CC_i$), LRMI($CC_i$)**

| | 2013 | 2014 | 2015 |
|---|---|---|---|
| $CC_i$ | 12,831,509 | 15,668,667 | 14,409,425 |
| M($CC_i$) | 1,779,935<br>(13.8%) | 2,722,425<br>(17.3%) | 2,724,591<br>(18.9%) |
| LRMI($CC_i$) | 95<br>(0.000053%) | 222<br>(0.000081%) | 319<br>(0.000117%) |

Figure 3 shows the top-20 PLDs and their particular LRMI term adoption. The figure illustrates that only a small amount of top-ranking providers utilise a range of different terms, while even among the top-20 providers only a small proportion uses more than 3 distinct LRMI terms.

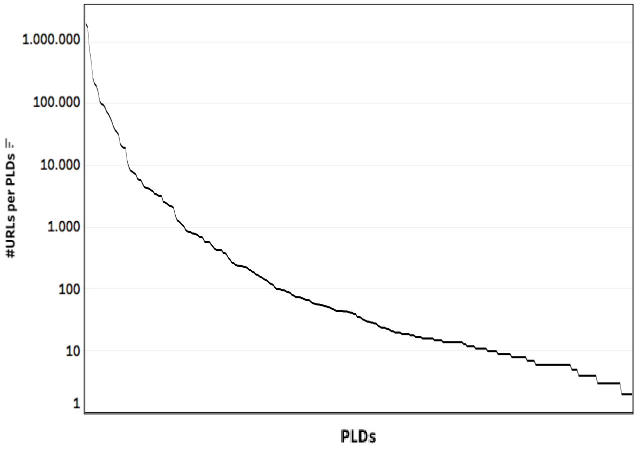

**Figure 2. Distribution across PLDs within LRMI($CC_{15}$)**

While the majority of PLDs indeed seems to be related to some form of learning, the relevance to learning varies heavily, with some more directly education-oriented websites such as *merlot.org* or *teacherspayteachers.com*, and some less relevant PLDs such as *ticketweb.com*.

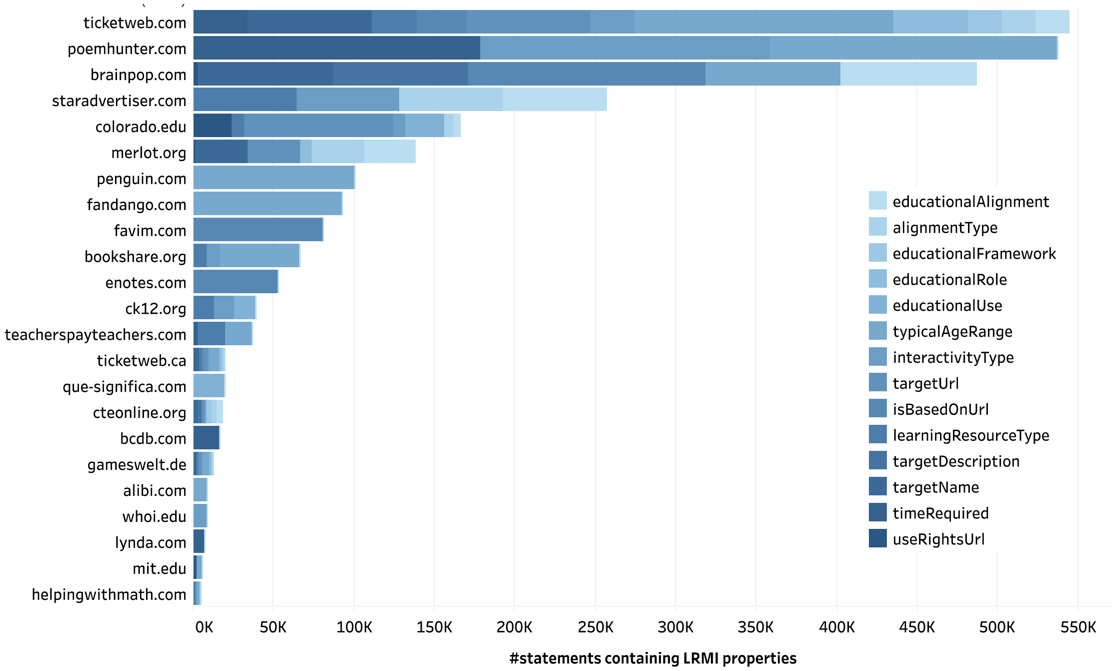

**Figure 3. LRMI property usage in LRMI($CC_{15}$) top-20 PLDs**

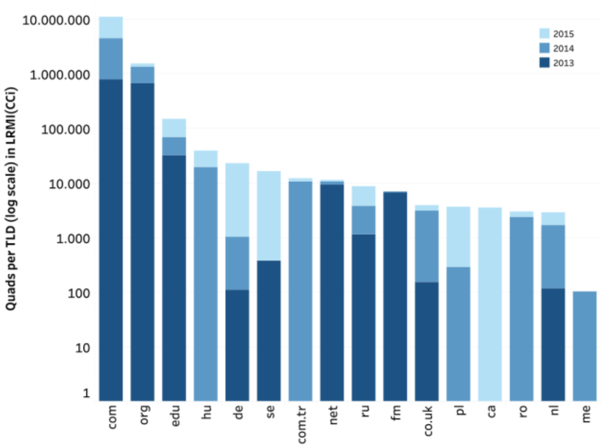

**Figure 4. Number of quads per TLD (top 15)**

Figure 4 depicts the number of quads per TLD (a more complete plot available online[13]), showing that *.com*, *.org* and *.edu* were indeed the most frequent TLDs in all three years. Plenty of new TLDs emerged since 2014, such *.as*, *.es*, *.fi*, *.gr*; *.eu*, while others started appearing in 2015, e.g. *.ca*, *.ch*, *.at*, *.jp*. This underlines the increasing diversity of LRMI providers. While some TLDs also disappeared from the crawl (e,g. *.su, .ie, .com.cn, .ac.uk, .fm*), this is partially due to the fact that erroneous statements were introduced over time, leading to their disappearance in *$LRMI(CC_{15})$* while being still partially present in *$LRMI'(CC_{15})$*.

Driven by the observation that different types of Websites/resource providers usually adopt different term combinations, we depict the co-occurrence graph of LRMI properties (incl. top-25 non LRMI properties) in *$LRMI(CC_{15})$* in Figure 5. Here, the size of a node indicates its degree of connectivity, while the thickness of the edge between two nodes indicates the frequency with which two nodes (properties) co-occur. For instance, *image* and *name* co-occur very frequently. The color-coding indicates sub-communities of commonly co-occurring sets of terms, detected through the *Louvain method* [14]. Here, typical combinations emerge, where for instance, the purple sub-graph indicates the strongly learning-related terms used by mostly learning-related PLDs, while other sub-graphs are less learning-related.

The network also illustrates the fact that particular terms such as *creator* have seen frequent use by learning resource (LRMI) providers and hence are part of the (purple) LRMI-specific sub graph, while particular LRMI terms, such as *isBasedOnUrl*, are frequently used in other contexts, probably due to their generic nature.

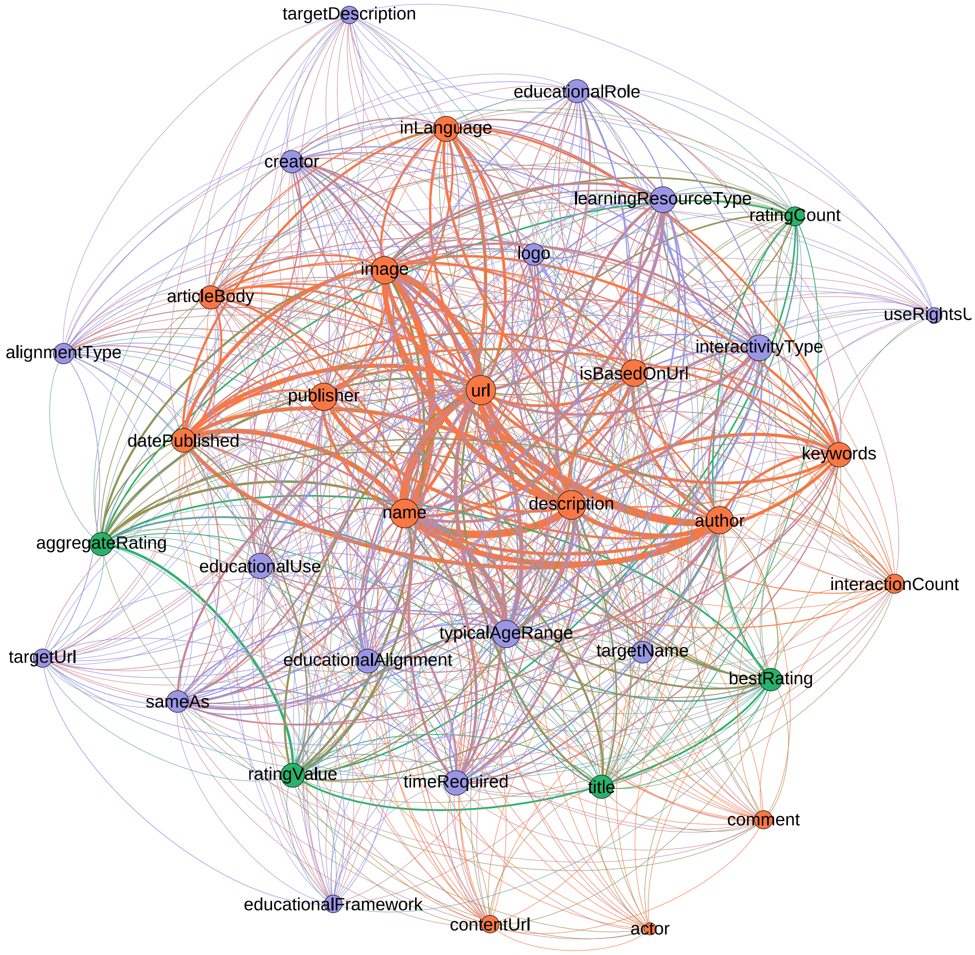

**Figure 5. Co-occurrence graph for LRMI properties and top-25 properties used in $LRMI(CC_{15})$**

Figure 6 presents a similar network visualisation, but limited to only LRMI properties. Given the central importance of *CreativeWork* and its subtypes, these were included too. The plot indicates that the particular combination of *typicalAgeRange*, *interactivityType* and *learningResourceType* is strongly connected and hence, highly representative for learning resource providers, and might constitute a particular pattern to look for when querying for strongly learning-related material.

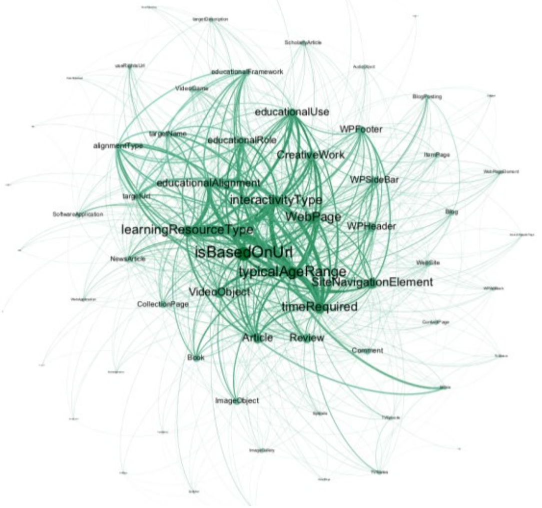

**Figure 6. Co-occurrence of LRMI properties in LRMI($CC_{15}$)**

# 6. LRMI MARKUP QUALITY

Quality of embedded markup varies heavily, requiring measures for data cleansing and improvement in order to enable reuse. We investigate the quality of markup annotations and introduce measures to improve data through applying a set of heuristics. In particular, we distinguish the following two types of observed errors: i) frequent errors and schema violations as discussed in [2], and ii) misuse of schema terms, i.e. the annotation of non-learning related resources through LRMI terms.

## 6.1 Common Errors

To address frequent errors, we implemented the heuristics proposed in [2] aimed at 1) fixing wrong namespaces, 2) handling undefined types and properties, 3) handling object properties misused as data type property, i.e. by assigning literal values. Note that all tables in this section show a limited set of ranks and examples, while more exhaustive tables are available online[13].

**Table 3. Common errors in LRMI'($CC_i$)**

| i | # Quads involving LRMI terms | # Quads in total | # affected PLDs |
|---|---|---|---|
| **Namespace errors** | | | |
| 2013 | 294 | 1,870 | 6 (6.06%) |
| 2014 | 1 | 2,128 | 7 (3.07%) |
| 2015 | 23,051 | 501,530 | 42 (11.17%) |
| **Undefined properties and types** | | | |
| 2013 | 73 | 61,231 | 36 (36.36%) |
| 2014 | 70 | 104,384 | 57 (25%) |
| 2015 | 953,527 | 1,172,893 | 137 (36.44%) |
| **Object properties used as datatype properties** | | | |
| 2013 | 64,475 | 596,226 | 66 (66.67%) |
| 2014 | 144,680 | 3342,115 | 143 (62.72%) |
| 2015 | 1,270,763 | 10,288,717 | 265 (70.48%) |

The number of corrected statements is shown in Table 3. The second column refers to quads involving LRMI terms, while the third column refers to all quads in the *LRMI'* corpora, i.e. even including non-LRMI statements co-occurring with LRMI statements. Most namespace issues seem due to typing errors, eg missing a missing a slash or "*https://*" instead of *http://*, while undefined terms often are caused by the misuse of upper/lower-cases in a case-sensitive context (Tables 5 and 6).

**Table 4. PLDs contributing most common errors in LRMI'($CC_i$) according to number of errors (left) and error rate (right)**

| Rank | Year | PLD | # Errors | % Errors |
|---|---|---|---|---|
| 1 | 2013 | merlot.org | 21,473 | 5,00 |
| | 2014 | bcdb.com | 44,325 | 1,40 |
| | 2015 | expedia.co.uk | 346,386 | 3,80 |
| 2 | 2013 | colorado.edu | 16,601 | 4,70 |
| | 2014 | colorado.edu | 18,948 | 4,70 |
| | 2015 | penguin.com | 337,145 | 10,50 |
| 3 | 2013 | mit.edu | 12,006 | 12,70 |
| | 2014 | merlot.org | 15,423 | 4,50 |
| | 2015 | expedia.com | 323,151 | 4,00 |
| 4 | 2013 | brainpop.com | 5319 | 9,00 |
| | 2014 | brainpop.com | 6973 | 10,00 |
| | 2015 | stanford.edu | 288250 | 88,10 |
| 5 | 2013 | curriki.org | 2326 | 6,20 |
| | 2014 | mit.edu | 6292 | 6,90 |
| | 2015 | expedia.ca | 94650 | 4,10 |

| Rank | Year | PLD | # Errors | % Errors |
|---|---|---|---|---|
| 1 | 2013 | bbc.co.uk | 1,752 | 75,00 |
| | 2014 | saraspublication.com | 14 | 41,20 |
| | 2015 | veeam.com | 2,465 | 96,30 |
| 2 | 2013 | geonetric.com | 10 | 31,30 |
| | 2014 | ditecinternational.com | 22 | 39,30 |
| | 2015 | ultracleantech.com | 819 | 92,60 |
| 3 | 2013 | pjjk.net | 12 | 27,30 |
| | 2014 | football-soccer-camps.com | 15 | 38,50 |
| | 2015 | teachersnotebook.com | 6,319 | 90,00 |
| 4 | 2013 | tlsbooks.com | 13 | 19,70 |
| | 2014 | weightlossnyc.com | 25 | 29,40 |
| | 2015 | stanford.edu | 288,250 | 88,10 |
| 5 | 2013 | davidfisco.com | 7 | 17,90 |
| | 2014 | timothylutheran.net | 8 | 27,60 |
| | 2015 | rubiksolve.com | 17 | 85,00 |

Compared to the whole markup corpus $M(CC_i)$, the statements in $LRMI'(CC_i)$ have a lower error rate. E.g. the wrong namespaces rate of $LRMI'(CC_{13})$ is 0.02% while the one for the $M(CC_{13})$ corpus is 1.23%. The rate of undefined properties/types is 0.66% compared to 5.82% in $M(CC_{13})$ as reported in [2]. The misuse of object properties is comparable in both corpora. The lower error rate in the *LRMI'* corpus presumably is due to the limited set of terms and the more constrained scope when following a specific vocabulary such as LRMI.

Investigating the PLDs which contribute the largest number respectively proportion of common errors (Table 4), it becomes apparent that providers with the highest error rates usually are part of the long tail of LRMI providers, often also using LRMI terms for non-learning related purposes.

**Table 5. Top-5 undefined properties in LRMI'(CCi)**

| Rank | Year | Property | # Quads | # PLDs |
|---|---|---|---|---|
| 1 | 2013 | useRightsUrl | 22067 | 9 |
| | 2014 | productioncompany | 30999 | 1 |
| | 2015 | isbasedonurl | 952889 | 37 |
| 2 | 2013 | offer | 10788 | 1 |
| | 2014 | useRightsUrl | 22582 | 6 |
| | 2015 | useRightsUrl | 28676 | 7 |
| 3 | 2013 | rating | 8091 | 1 |
| | 2014 | alternatename | 13325 | 1 |
| | 2015 | intendedEndUserRole | 8434 | 3 |
| 4 | 2013 | intendedEndUserRole | 6696 | 6 |
| | 2014 | offer | 8539 | 1 |
| | 2015 | embedURL | 7440 | 5 |
| 5 | 2013 | inlanguage | 2659 | 1 |
| | 2014 | intendedEndUserRole | 7045 | 3 |
| | 2015 | company | 6871 | 1 |

The most frequently observed undefined properties (types) for all years under observation are shown in Table 5 (Table 6). A large proportion seems due to mistyping of established schema.org types, while others lack any obvious relation with existing terms.

**Table 6. Top-5 undefined types in LRMI'(CCi)**

| Rank | Year | Type | # Quads | # PLDs |
|---|---|---|---|---|
| 1 | 2013 | EducationalEvent | 6004 | 1 |
| | 2014 | EducationalEvent | 3047 | 1 |
| | 2015 | offer | 100516 | 1 |
| 2 | 2013 | UserComment | 20 | 1 |
| | 2014 | Therapist | 25 | 1 |
| | 2015 | headline | 6724 | 1 |
| 3 | 2013 | CompetencyObject | 4 | 1 |
| | 2014 | UserComment | 23 | 1 |
| | 2015 | URL | 693 | 1 |
| 4 | 2013 | Webpage | 2 | 1 |
| | 2014 | learningResourceType | 21 | 1 |
| | 2015 | webpage | 360 | 1 |
| 5 | 2013 | about | 1 | 1 |
| | 2014 | EducationalEvent | 19 | 1 |
| | 2015 | musicrecording | 296 | 1 |

As apparent in Table 6, type errors usually occur within one particular PLD only, what indicates that frequency-based signals provide useful hints when filtering PLDs or markup data in general.

**Table 7. Erroneous/fixed quads, docs and PLDs**

| | 2013 | 2014 | 2015 |
|---|---|---|---|
| # quads | 520,815 (5.63%) | 1,601,796 (6.10%) | 6,179,097 (8.84%) |
| # docs | 46,382 (55.15%) | 369,772 (85.81%) | 754,863 (81.21%) |
| # PLDs | 75 (75.76%) | 154 (67.54%) | 291 (77.39%) |

The total and relative amount of documents, quads and PLDs which contained common errors in $LRMI'(CC_i)$ and were fixed through applying the heuristics mentioned above are shown in Table 7. Numbers in brackets indicate the relative amounts compared to the entire dataset. As shown, while the number of affected quads is comparably low, approximately 70% of all PLDs and more than 80% of documents in 2014 and 2015 contain incorrect statements, which are fixed by applying the aforementioned heuristics. We also observe a trend of increasing amounts of erroneous statements from 2013 to 2015. This underlines the need for additional data processing before reusing or interpreting markup data.

## 6.2 Misuse of Vocabulary Terms

Another challenge when interpreting and reusing markup data is the often ambiguous use of properties, caused by varying interpretations of term semantics. In the LRMI case, one can observe a large amount of documents that contain LRMI annotations which appear to be not learning-related. For instance, the property *typicalAgeRange* is often used by Websites which provide adult content. Considering the original LRMI specification which defines this property as an attribute to indicate the educational suitability of a particular learning resource, this constitutes an unintended use. A more thorough discussion can be found in Section 7. While annotations of such kind are problematic when reusing and recommending learning resources, we apply data filtering based on a domain blacklist[14] which contains 1,078,273 adult content domains. We filter out all the quads that originate from these domains. The amount of quads and documents that were filtered in $LRMI'(CC_i)$ based on the domain blacklist is reported in Table 8. The fourth row (*subdomains*) indicates subdomains from a PLD (*lap.hu*) which was only partially filtered, and hence not included into the #PLDs.

**Table 8. Number of filtered quads, docs, and PLDs**

| | 2013 | 2014 | 2015 |
|---|---|---|---|
| # quads | 88,829 (0.96%) | 38,376 (0.15%) | 36,538 (0.05%) |
| # docs | 1,594 (1.9%) | 576 (0.13%) | 525 (0.06%) |
| # PLDs | 8 (8.08%) | 27 (11.84%) | 23 (6.12%) |
| # subdomains | 0 | 8 (0.44%) | 11 (0.27%) |

According to a manual evaluation of the filtered PLDs, this processing step filtered adult content with a recall of 96% and hence, helped improving the suitability of LRMI data. While this processing step addressed one of the most obvious issues emerging from diverse interpretations and usage of schema terms, which are more deeply discussed in Section 7, it is important to note that we observe a wide range of content (provider) types, where often it is debatable whether or not a particular resource (or Website) is considered a resource of relevance for learning.

[14] http://dsi.ut-capitole.fr/blacklists/index_en.php

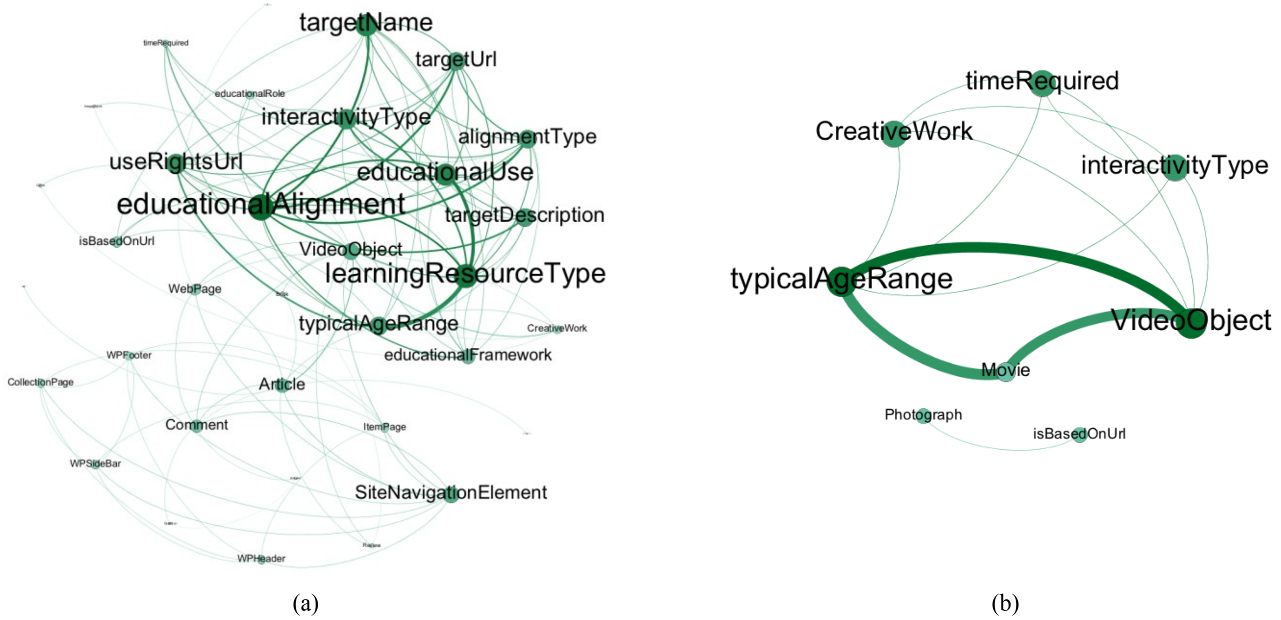

**Figure 7. Term co-occurrence graphs for (a) top-k learning-related PLDs and (b) filtered adult content PLDs**

Figure 7 investigates the term co-occurrence of such filtered PLDs (Figure 7), where the graph on the right (b) shows the term co-occurrence within the data extracted from the filtered adult content PLDs (*n=24, i=2015*), while (a) shows the term co-occurrence for the top 24 PLDs with strong relevance to learning-related content in the same year. It becomes apparent that term usage and distribution strongly varies dependent on the scope of the content, and hence, constitute useful features for filtering, clustering or classifying LRMI data from different sources and of different nature. Current experiments with unsupervised models (K-Means, LDA) support this finding and suggest potential for detecting different term interpretations. As part of related work on fusing entity-centric markup facts [9], we have already demonstrated that the exploitation of a range of features of the underlying PLDs as well as the provided markup data leads to strong results when aiming at detecting correct and diverse data.

# 7. DISCUSSION

Here we discuss the key findings of our work as well as the limitations of our approach.

## 7.1 Key Findings

In this section, we reflect on the observations presented in this study, in particular with the aim to aid the reuse and interpretation of LRMI data by data consumers, such as Web search or recommendation engines, and to identify future directions for the ongoing refinement of LRMI terms as part of the DCMI task group on LRMI[15]. Key findings are summarised below:

**I. Power-law distribution of LRMI markup.** Few providers (PLDs) contribute vast amounts of data (Section 5), where the top 10% of contributors provide 98.4% of all quads in *LRMI($CC_{15}$)*. While this mirrors observations for markup in general [10], it provides clues for data consumers aiming at efficient means to frequently crawl and extract LRMI-specific data, for instance, when dynamically creating knowledge bases or indexes of learning resources from LRMI markup. Here, the application of highly focused crawling strategies targeting the most probable data providers seems to be an efficient means to obtain available markup data.

**II. Frequent errors.** Although LRMI markup contains fewer errors than markup in general (Section 6), vast amounts of erroneous statements can be observed, where approximately 80% of all documents and PLDs contribute one or more incorrect statements (2015), even when assessing only the most frequent issues. The general trend indicates rising rates of erroneous statements (2013-2015). This calls for the application of data cleansing and improvement strategies when reusing and interpreting markup data, where simple heuristics already yield significant improvement in overall data quality. Furthermore, the fact that widely misused terms and properties are usually used by a very small amount of PLDs indicates that frequency-based features provide indeed strong indicators when aiming to filter or fuse data from markup [7]. In addition, the observed undefined schema terms and types (Tables 5 and 6) can inform future discussions about the extension of schema.org, as a means for bootstrapping term recommendations.

**III. Biased term adoption pattern.** The adoption of LRMI terms strongly differs across terms (Section 4) and appears to depend on a variety of characteristics, with simple and generic properties appearing widely popular while increasing complexity and specificity of terms correlates with limited adoption. In particular, there is a strong tendency for using datatype properties, or, misusing any property as such, with more than 96% of statements referring to literals as objects (rather than URIs) in 2015. This has strong implications for the required processing (data consumer side) as well as future vocabulary developments.

**IV. Unintended use of terms and types.** As underlined by Section 6, terms are often applied in contexts not intended or originally foreseen, for instance, the use of LRMI terms to describe adult content. While this is not necessarily considered a schema violation [10], it also leads to the need for further processing to unambiguously interpret and reuse markup data. For instance, the mere use of LRMI terms does not provide accurate

[15] http://wiki.dublincore.org/index.php/AB-Comm/ed/LRMI/TG, co-chaired by one of the authors

indicators of whether or not a particular creative work carries an inherent learning value. While simple heuristics and filtering steps appear to be successful for filtering (Section 6), more sophisticated means are required to better cluster and classify resources and resource providers (PLDs).

Observations II-IV underline the heterogeneous and largely unstructured nature of markup data, raising the need for tailored mapping and fusion approaches to address issues such as identity resolution and incorrectly annotated data when consuming and interpreting markup data.

From a vocabulary design perspective, several observations will be considered for future developments, for instance, the application of more specific labels or descriptions to frequently misused terms, such as *timeRequired*, or *typicalAgeRange*. In addition, the strong tendency towards flat entity descriptions and the lack of acceptance for embedding actual graphs, that is object-object relationships, into Web pages particularly impacts the *educationalAlignment* property, which is seen as a core element of LRMI, enabling to associate any resource with a particular educational framework through the *AlignmentObject*. While this modeling approach has seen only very limited adoption (Figure 1), this problem is elevated by the improper use of *AlignmentObject*, often failing to provide an *educationalFramework* or *targetUrl*. This can be explained with the apparent tendency towards simple datatype statements as well as the fact, that learning resources are not necessarily tied to formal educational frameworks. However, given that this constitutes the primary method of marking a creative work as a learning resource, this observation raises the need to expand the LRMI specification towards a wider range of cases and simpler means to associate learning objectives with resources.

## 7.2 Limitations

While we have extracted a first corpus for studying the adoption of LRMI and provided an initial set of findings, we also like to discuss limitations of this work. In particular, with respect to the dataset, we have exploited the Common Crawl as the largest publicly available general-purpose crawl under the assumption that it represents a representative sample of the Web at a given point in time. However, the nature of the Common Crawl leads to a number of constraints regarding the interpretation of the data. The varied scope and scale of the crawl iterations limit the generalisability of our findings concerning trends and evolution of LRMI data over time. In this respect, findings might be impacted by biased crawl iterations, where, for instance, in one year a larger proportion of potential LRMI providers might be crawled than in others. In addition, general conclusions about LRMI adoption (or lack thereof) are hard to draw given that the crawl is not set up to capture a representative sample of LRMI providers in particular. This could be alleviated by iterative focused crawls using a consistent set of crawling seeds, where future work is concerned with extracting and analysing LRMI data from a targeted crawl of PLDs of potential LRMI-relevance, for instance, educational organisations, learning material providers or libraries.

We also would like to emphasise deviations in the dataset sizes provided on our paper-related website[13]. While the data dumps contain quads of the form *{s, p, o, u}*, the corresponding RDF datasets were generated through a transformation process, where for each entity description consisting of a set of triples *{s, p, o}*, a separate statement was added, which relates *s* to a particular document URL *u*, what leads to a slight increase of the dataset size compared to the original amount of quads. All sizes in the document refer to the original quads rather than the datasets after triplification. An additional effect of the RDF transformation is the removal of duplicate triples occurring on the same document. While this seems reasonable, for instance, when building recommender systems, the frequency with which a particular triple occurs within and across documents or PLDs provides important signals when attempting to fuse or filter facts [9].

In addition, it is worth highlighting that the cleansing and filtering steps in Section 6 only provide an initial set of rather pragmatic processing steps aimed at understanding and improving the quality of LRMI data. Further processing is required to better categorise, filter and interpret data. For instance, the observation that the term distribution of a PLD provides signals about its general scope suggests that clustering techniques can be applied to separate strictly learning-related PLDs from other, less LRMI-specific content and providers.

# 8. CONCLUSION & OUTLOOK

We have assessed the adoption of LRMI vocabulary terms on the Web and provided the yet largest available corpus of LRMI markup crawled from the Web. While a significant amount of Web pages in the Common Crawl contain embedded markup, namely 30.5% of more than 1.7 billion documents in 2015, the proportion of documents with LRMI statements is comparably small. However, the total amount of quads including or co-occurring with LRMI statements adds up to 105,359,363 (44,108,511) in 2016 (2015), showing an increase by 139% (51%) from 2015 (2014). Given the still very recent nature of the LRMI vocabulary and its continuous evolution, its increasing adoption suggests potential for exploiting such data as part of recommender systems, search engines or to dynamically populate knowledge graphs of learning-related resources and entities. Since this study has exploited a general-purpose Web crawl as a representative sample of the Web, a more focused crawl of educational and learning-related sites is likely to obtain LRMI markup in even higher quantity and quality. In addition, it is also worthwhile to note that a variety of non-LRMI terms (e.g., *CollegeOrUniversity*, *EducationalOrganization*) is used for the annotation of educational and learning-related entities.

Errors are less frequent than in general markup data but still increasingly prevalent (Section 6). Hence, significant processing is required when using and interpreting LRMI markup. We apply simple processing steps aimed at correcting frequent errors and to filter out erroneous statements. The dynamic nature of Web documents and embedded entity markup suggests a strong potential for creating dynamic and focused knowledge graphs through frequently crawling, extracting and consolidating entity markup from the Web. In this context, we are currently investigating data fusion techniques tailored to the specific needs of Web markup [9], with the aim to complement existing knowledge bases and linked data in general, as well as learning resource metadata and related entity-centric knowledge in particular.

# 9. ACKNOWLEDGMENTS

This work has been partially supported by the H2020 programme of the European Union under grant agreement No 687916 – AFEL project (http://afel-project.eu/) and the COST Action KEYSTONE (IC1302).

# 10. REFERENCES

[1] Meusel R., Petrovski P., and Bizer C. 2014. The WebDataCommons Microdata, RDFa and Microformat Dataset Series. In Proc. of the 13th International Semantic Web Conference (ISWC14), Springer-Verlag New York, Inc., New York, NY, USA, 277-292.

[2] Meusel R., Paulheim H. 2015. Heuristics for fixing common errors in deployed schema.org microdata. In Proc. of the ESWC 2015 Conference - The Semantic Web. Latest Advances and New Domains. Springer, 2015. 152–168.

[3] Buil Aranda, C., Hogan, A., Umbrich, J., Vandenbussche, P. Y., SPARQL Web-Querying Infrastructure: Ready for Action?, The 12th International Semantic Web Conference (ISWC2013).

[4] Guha, R. V., Brickley, D., Macbeth, S., Schema.org: evolution of structured data on the web. Commun. ACM 59, 2, 44-51. http://dx.doi.org/10.1145/2844544.

[5] d'Aquin, M., Adamou, A., Dietze, S. 2013. Assessing the Educational Linked Data Landscape. In Proceedings of ACM Web Science 2013 (WebSci2013), Paris, France, May 2013.

[6] Yu, R., Fetahu, B., Gadiraju, U., Dietze, S., A Survey on Challenges in Web Markup Data for Entity Retrieval, Poster paper at 15th International Semantic Web Conference (ISWC2016), Kobe, Japan, October 2016.

[7] Yu, R., Gadiraju, U., Zhu, X., Fetahu, B., Dietze, S., Entity summarisation on structured web markup. In The Semantic Web: ESWC 2016 Satellite Events. Springer, 2016.

[8] Dietze S., Yu H. Q., Giordano D., Kaldoudi E., Dovrolis N., Taibi D. 2012. Linked Education: interlinking educational Resources and the Web of Data. ACM Symposium On Applied Computing (SAC-2012), Special Track on Semantic Web and Applications.

[9] Yu, R., Gadiraju, U., Fetahu, B., S. Dietze, Query-centric Data Fusion on Structured Web Markup. In IEEE 33rd International Conference on Data Engineering (ICDE2017). IEEE, 2017.

[10] Dietze S., Sanchez-Alonso S., Ebner H., Yu H. Q., Giordano D., Marenzi I., Pereira Nunes B. 2013. Interlinking educational resources and the web of data: a survey of challenges and approaches. Emerald Program: electronic library and information systems, 47(1), 60-91.

[11] Hogan, A., Hitzler, P., Krzysztof, J., Linked Dataset description papers at the Semantic Web journal: A critical assessment, Semantic Web Journal, Vol. 7, No. 2, 2016.

[12] De Santiago, R. and Raabe, A.L.A. (2010), "Architecture for Learning Objects Sharing among Learning Institutions-LOP2P", IEEE Transactions on Learning Technologies, April-June, 2010, pp. 91-5.

[13] Taibi, D., Dietze, S., Towards embedded markup of learning resources on the Web: a quantitative Analysis of LRMI Terms Usage, in Companion Publication of the IW3C2 WWW 2016 Conference, IW3C2 2016, Montreal, Canada, April 11, 2016.

[14] Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R. & Lefebvre, E. (2008), Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008.