However, several explanations have to be taken into account here, in order to put the results into perspective. First of all, since we currently use the CreativeWork-specific WDC subset, all subtypes of Creative Works are not considered in this study. Hence, a drop in quads in our data might as well be caused by certain key providers (see next Section) adopting a more fine-grain annotation strategy, preferring more specific types, such as s:Article or s:Book rather than the generic Creative Work type.

Another investigation is the lack of consistency between Common Crawls over the years, where a URL (or document) crawled in 2013 is not necessarily part of the 2014 crawl, despite that being the case for the majority of documents. As shown in the following section, for some key LRMI providers, the amount of documents overall in our investigated dataset has dropped significantly.

To understand the nature of annotated works, we also report the learning resource types indicated explicitly through the *learningResourceType* predicate (Figure 2). These include "Worksheet" (11.6% in 2013 and 12.2% in 2014), "Games" (9% in 2013 and 8.7% in 2014), Assessment (7.3% in 2013 and 7.5% in 2014), "PowerPoint presentation" (6.4% in 2013 and 6% in 2014) and "Quiz" (2.5% in 2013 and 2.3% in 2014). For a discussion of the Null values, please see Section 5.

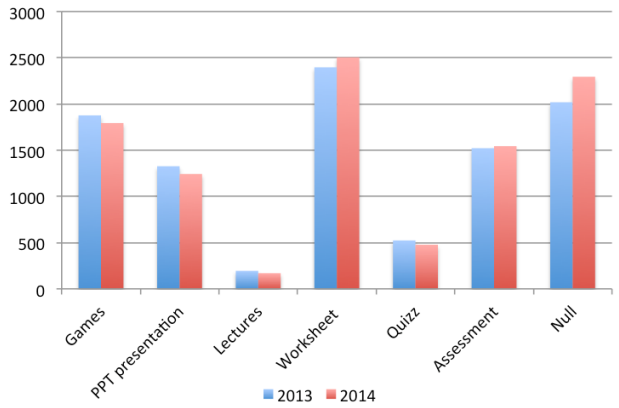

Figure 2: Main learning resource types

4. DISTRIBUTION ACROSS PLDS

In the class-specific subset under investigation the total number of PLDs using LRMI properties in 2013 is 21, while in 2014 this number increase to 33, thus confirming a positive trend in the diffusion of LRMI properties amongst pay level domains (PLDs). Figure 3 provides an overview of the distribution of markup per PLD. For this figure all the LRMI properties related to the three classes *CreativeWork*, *AlignmentObject* and *EducationalAudience* have been taken into consideration.

As shown, a small number of PLDs contains the majority of markup related to LRMI properties. However, even though 5 PLDs include 99% of LRMI properties there are others PLDs that include relevant learning materials such as: teachersnotebook.com, senteacher.org, pomagalo.com, thegateway.org and bbc.co.uk.

The claz.org web site appears in the list due to the frequent use of the property *isBasedOnUrl*, even if it is not a website specialized in educational content.

More details about the properties used by the main PLDs related to educational content are reported on Table 4.

The analysis of WhoIs records of the PLDs has revealed that in 2013 the majority of PLDs is registered in the US (12) and the UK (5). While in 2014, PLDs registered in US and UK are 18 and 7, in addition a more diverse set of countries such as Brasil, France, Russia, Latvija are also represented (Table 2).

Table 2: PLD registration

2013		2014	
US	12	US	18
UK	5	UK	7
Russia	1	Russia	1
Italy	1	Italy	1
Bulgaria	1	Brasil	2
Algeria	1	Netherlands	2
		France	1
		Latvia	1

On further inspection, it appears that a number of PLDs are using LRMI statements for non-intended purposes. In 2013 and 2014 we detected respectively 8 and 12 PLDs not related to education. Examples of PDLs using LRMI properties for content not strictly related to educational are: "oneunlock.com", "orbussoftware.com", "cartoni-animati.org" or "cmonbook.com". However, the number of quads extracted for these PLDs is comparably low. In particular, the properties *isBasedOnUrl* and *timeRequired* seem used (or misused) by these PLDs.

Table 3 compares the total number of documents in our dataset (including also documents not containing LRMI properties) with the number of quads containing LRMI terms.

Table 3: Total number of documents and number of quads using LRMI properties across PLDs

PLD	2013		2014	
	#quads	#docs	#quads	#docs
brainpop.com	513090	15062	559992	6829
claz.org	439102	24434	469600	23850
merlot.org	218466	10641	171770	8464
teacherspayteachers.com	55755	18322	52995	17216
curriki.org	11444	322	13115	376
teachersnotebook.com	-	-	550	272
thegateway.org	668	66	-	-
pomagalo.com	118	38	216	68
senteacher.org	84	20	264	28

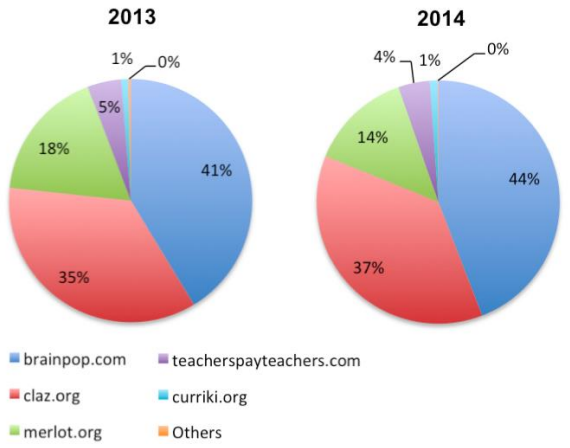

Figure 3: Distribution of LRMI quads in 2013 and 2014 amongst PLDs

Table 3 provides useful insights to understand the shape of our dataset and the evolution of quads from 2013 to 2014. In fact, for the most representative PLDs the total numbers of documents included in the dump decreased between the two years. In the case of brainpop.com even if the number of crawled documents has decreased by more than 50%, the number of quads including LRMI properties increased. In other cases, the reduction of crawled documents has led to a drastic reduction of the quads related to LRMI properties.

These decreasing number of documents might be explained by the inconsistency of the Common Crawl over the years (as described in the previous section) or the adoption of more specific subtypes by the Website provider, what would lead to the documents not showing up in our type-specific subset.

5. OBSERVED ERRORS IN LRMI STATEMENTS

The numbers reported in Table 1 paint a promising picture of the use of LRMI properties. However, the analysis of the actual statements, i.e. the values associated with these properties reveals significant issues. While a number of predicates seem to be used in a meaningful way, for instance, the *typicalAgeRange* property is involved in seemingly correct statements and only shows null values for 2% of the statements in both years, other predicates seem to be not used in the correct and intended way. Specifically, the occurrence of null values is dominant in some cases.

Regarding the *learningResourceType* property the percentage of null values is very low (9.7% in 2013 and 11.2% in 2014), but still reasonable high if compared with other values.

The observed values for the *educationalUse* property in 2014 reveal that Null values are the most used (93%). The analysis of the values undertaken for other properties reveal a similar state. The values for the *interactivityType* properties in 2014 are divided as follows: 96.8% Null, 3.1% active, 0.1% mixed.

In WDC2013 and WDC2014, the *timeRequired* property has been valorized with zero in 80.5% and 80.1% of the RDF quads.

The recommended value for the *alignmentType* property are: 'assessed', 'teaches', 'requires', 'textComplexity', 'readingLevel', 'educationalSubject', and 'educationLevel', but the analysis of the data has revelead that only 'educationalSubject' has been used in both years.

Moreover, some frequent errors related to schema violations have been observed:

- the *typicalAgeRange* property reported in the table in both years often (77% in 2013 and 80% in 2014) refers to instances of the class *AlignmentObject*, while the valid range is defined as instances of the *CreativeWork* class only.
- We detected capitalization errors (e.g. EducationalUse and educationaluse respectively in 5 and 3 quads) for the *educationalUse* property in 2014, while no errors detected in the 2013 collection.
- The *alignmentType* where the valid domain is instances of the *AlignmentObject* class, also is used (only in 4 quads, though) with the *Schema.org/CompetencyObject* which is not defined as part of Schema.org.

Table 4: The main PLDs related to educational content with LRMI property markup

	brainpop.com		merlot.org		teacherspayteachers.com		curriki.org		pomagalo.com		senteacher.org		thegateway.org*	artnc.org
	2013	**2014**	**2013**	**2014**	**2013**	**2014**	**2013**	**2014**	**2013**	**2014**	**2013**	**2014**	**2013**	**2014**
educationalAlignment	83975	97046	51276	40276									167	3
educationalUse							2011	2282	38	68	28	88		3
timeRequired					18585	17665			4	12				1
typicalAgeRange	83975	97046			18585	17665	5402	6258			28	88		4
interactivityType							2011	2282	38	68				2
learningResourceType					18585	17665	2011	2282	38	68	28	88		2
isBasedOnUrl	163300	162035												2
useRightsURL	13890	9773	2585	2126			9	11						1
alignmentType			51276	40276									167	
targetDescription	83975	97046											167	
targetName	83975	97046	51276	40276										
targetUrl			51276	40276									167	
educationalRole			10777	8540										

* The PDL http://thegateway.org is present only in 2013 since it has been closed.

6. CONCLUSIONS

In this study, we have assessed the adoption of LRMI vocabulary terms on the Web. While a significant amount of Web pages (2.01 billion pages) and PLDs (2.72 million) in the Common Crawl contain embedded markup, the proportion of LRMI statements is comparably small. However, as our current investigation was limited to the CreativeWork subset of the WDC, this approach did not consider any CreativeWork subtypes, potentially missing a significant amount of LRMI data. Our study also finds that a large proportion of statements are of limited usage so far. With respect to growth, within the scope of the Common Crawl, minor growth of LRMI statements is detected (2.15% percent increase) from 2013 to 2014. While some terms even have seen a drop in adoption, this might be explained with the variance of the crawled URLs between both years. A more controlled study of continuously recrawling a focused set of URLs for a longer period of time would help in further investigating the evolution.In addition, it is also worthwhile to note that learning-related resources are annotated with a number of non-LRMI terms from the schema.org vocabulary, for instance, *CollegeOrUniversity*, *EducationalOrganization*, *School*, *Museum*, *Article*, *Book*.

On the other hand, significant growth has been detected by the number of LRMI adopters (PLDs) over time, which increased by nearly 50% from 2013 to 2014. Therefore, the current investigation suggests that a targeted crawl of potential LRMI providers would surface a significant amount of embedded markup that will emerge into an unprecedented source of knowledge about educational resources on the Web. Spreading awareness about LRMI and its use seems among the key aims of current working groups such as the LRMI DCMI Task Force and related W3C Community Groups.

7. 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 COST Action KEYSTONE (IC1302).

8. 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 (ISWC '14), Mika P., Tudorache T., Bernstein A., Welty C., Knoblock C., Vrandeăić D., Groth P., Noy N., Janowicz K., and Goble C. (Eds.). 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] 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.

[4] Fetahu, B., Dietze, S., Nunes, B. P., Casanova, Taibi, D., M. A., Nejdl, W. 2014. A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. *In Proceedings of 11th Extended Semantic Web Conference*

[5] 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*.

[6] 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. doi: 10.1108/00330331211296312.

[7] Taibi, D., Dietze, S., Fetahu, B., Fulantelli, G. 2014. Exploring type-specific topic profiles of datasets: a demo for educational linked data, in Poster & System Demonstration Proceedings of 13th International Semantic Web Conference (ISWC2014), Riva Del Garda, Italy, October 2014.