Towards embedded Markup of Learning Resources on the Web: an Initial Quantitative Analysis of LRMI Terms Usage

Davide Taibi
National Research Council of Italy
Institute for Educational Technologies

Stefan Dietze
L3S Research Center, Germany
Educational Linked Data

Initiatives

- W3C Library Linked Data Incubator Group
- Linked Library Data group on DataHub
- LinkedUniversities.org
- LinkedEducation.org
- W3C Linked Open Education Community Group
- ...
The Web as a knowledge base: semantics on the Web?

- The Web: approx. 46,000,000,000,000 (46 trillion) Web pages indexed by Google

vs

- Linked Data: approx. 1000 datasets & 100 billion statements
  - different order of magnitude wrt scale & dynamics

- Other „semantics“ (structured facts) on the Web?
Embedded semantics: Web page markup & schema.org

- Embedded markup (RDFa, Microdata, Microformats) for interpretation of Web documents (search, retrieval)
- Arbitrary vocabularies; schema.org used at scale: (700 classes, 1000 predicates)
- Adoption on the Web: 26% (2014 Google study of 12 bn Web pages)
- “Web Data Commons” (Meusel & Paulheim [ISWC2014])
  - Markup from Common Crawl (2.2 billion pages): 17 billion RDF quads
  - Markup in 26% of pages, 14% of PLDs in 2013 (increase from 6% in 2011)
- Same order of magnitude as “the Web”

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Forrest Gump</h1>
  <span>Actor: <span itemprop="actor">Tom Hanks</span></span>
  <span itemprop="genre">Drama</span>
  ...
</div>
```

<table>
<thead>
<tr>
<th>RDF statements</th>
</tr>
</thead>
<tbody>
<tr>
<td>node1</td>
</tr>
<tr>
<td>node1</td>
</tr>
<tr>
<td>node1</td>
</tr>
<tr>
<td>node2</td>
</tr>
<tr>
<td>node2</td>
</tr>
<tr>
<td>node3</td>
</tr>
<tr>
<td>node3</td>
</tr>
</tbody>
</table>
Other “semantics“ (structured facts) on the Web!

Knowledge graphs and LD
(Yago, Freebase, Pubmed, DBLP etc)
Learning Resources Metadata Initiative

http://www.lrmi.net

- LRMI specification: a collection of properties to describe educational resources.

- LRMI specification added to Schema.org in April 2013

<table>
<thead>
<tr>
<th>CreativeWork</th>
<th>AlignmentObject</th>
<th>EducationalAudience</th>
</tr>
</thead>
<tbody>
<tr>
<td>educationalAlignment</td>
<td>alignmentType</td>
<td>educationalRole</td>
</tr>
<tr>
<td>educationalUse</td>
<td>educationalFramework</td>
<td></td>
</tr>
<tr>
<td>timeRequired</td>
<td>targetDescription</td>
<td></td>
</tr>
<tr>
<td>typicalAgeRange</td>
<td>targetName</td>
<td></td>
</tr>
<tr>
<td>interactivityType</td>
<td>targetUrl</td>
<td></td>
</tr>
<tr>
<td>learningResourceType</td>
<td></td>
<td></td>
</tr>
<tr>
<td>isBasedOnUrl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Method and Research Questions

Research Questions:

• Evolution of LRMI adoption over time
• Most represented Learning Resource Type
• Distribution of LRMI terms across PLDs
• Observed frequent errors in LRMI statements

Dataset

• Web Data Commons (webdatacommons.org) Common Crawl
• Common Crawl web corpus November 2013 and December 2014
### Dataset

**CreativeWork subset**

<table>
<thead>
<tr>
<th>Year</th>
<th>quads</th>
<th>entities</th>
<th>documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>51,601,969</td>
<td>10,469,565</td>
<td>11,681,807</td>
</tr>
<tr>
<td>2014</td>
<td>50,901,532</td>
<td>11,861,807</td>
<td>4,343,951</td>
</tr>
</tbody>
</table>

**CreativeWork subset containing LRMI properties**

<table>
<thead>
<tr>
<th>Year</th>
<th>quads</th>
<th>entities</th>
<th>documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>1,242,094</td>
<td>949,057</td>
<td>151,657</td>
</tr>
<tr>
<td>2014</td>
<td>1,268,951</td>
<td>972,542</td>
<td>143,884</td>
</tr>
</tbody>
</table>

**Complete LRMI subset**

<table>
<thead>
<tr>
<th>Year</th>
<th>quads</th>
<th>entities</th>
<th>documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>10,636,873</td>
<td>1,461,093</td>
<td>83,791</td>
</tr>
<tr>
<td>2014</td>
<td>30,599,024</td>
<td>4,182,541</td>
<td>430,861</td>
</tr>
</tbody>
</table>
Evolution of LRMI adoption over time

Less documents but dense

Fine-grain annotation
- 89 distinct classes in 2013
- 157 distinct classes in 2014
# Evolution of LRMI adoption over time

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#docs</td>
<td>#entities</td>
<td>#quads</td>
<td>#docs</td>
<td>#entities</td>
<td>#quads</td>
</tr>
<tr>
<td>Class: CreativeWork</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>educationalAlignment</td>
<td>12710</td>
<td>13286</td>
<td>135418</td>
<td>11335</td>
<td>11621</td>
<td>137325</td>
</tr>
<tr>
<td>educationalUse</td>
<td>381</td>
<td>391</td>
<td>2079</td>
<td>478</td>
<td>541</td>
<td>2450</td>
</tr>
<tr>
<td>timeRequired</td>
<td>18704</td>
<td>18594</td>
<td>18594</td>
<td>17981</td>
<td>17957</td>
<td>17957</td>
</tr>
<tr>
<td>typicalAgeRange</td>
<td>20982</td>
<td>103339</td>
<td>108417</td>
<td>20525</td>
<td>115310</td>
<td>121195</td>
</tr>
<tr>
<td>interactivityType</td>
<td>359</td>
<td>361</td>
<td>2049</td>
<td>447</td>
<td>450</td>
<td>2357</td>
</tr>
<tr>
<td>learningResourceType</td>
<td>18704</td>
<td>18977</td>
<td>20665</td>
<td>17981</td>
<td>18581</td>
<td>20488</td>
</tr>
<tr>
<td>isBasedOnUrl</td>
<td>29432</td>
<td>444559</td>
<td>605321</td>
<td>26613</td>
<td>472714</td>
<td>631762</td>
</tr>
<tr>
<td><strong>useRightsUrl</strong></td>
<td>14311</td>
<td>16490</td>
<td>16491</td>
<td>8932</td>
<td>11911</td>
<td>11911</td>
</tr>
<tr>
<td>Class: AlignmentObject</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alignmentType</td>
<td>10653</td>
<td>51447</td>
<td>51447</td>
<td>8460</td>
<td>40276</td>
<td>40276</td>
</tr>
<tr>
<td>targetDescription</td>
<td>2077</td>
<td>84142</td>
<td>84142</td>
<td>2876</td>
<td>97046</td>
<td>97046</td>
</tr>
<tr>
<td>targetName</td>
<td>2059</td>
<td>135251</td>
<td>135251</td>
<td>11334</td>
<td>137319</td>
<td>137322</td>
</tr>
<tr>
<td>targetUrl</td>
<td>10651</td>
<td>51443</td>
<td>51443</td>
<td>8458</td>
<td>40276</td>
<td>40276</td>
</tr>
<tr>
<td>educationalFramework</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Class: EducationalAudience</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>educationalRole</td>
<td>10634</td>
<td>10777</td>
<td>10777</td>
<td>8464</td>
<td>8540</td>
<td>8540</td>
</tr>
</tbody>
</table>

*Highlighted properties with changes.*
Most represented Learning Resource Type

<table>
<thead>
<tr>
<th>Learning Resource Type</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Worksheet</td>
<td>11.6%</td>
<td>12.2%</td>
</tr>
<tr>
<td>Games</td>
<td>9%</td>
<td>8.7%</td>
</tr>
<tr>
<td>Assessment</td>
<td>7.3%</td>
<td>7.5%</td>
</tr>
<tr>
<td>PPT presentation</td>
<td>6.4%</td>
<td>6%</td>
</tr>
<tr>
<td>Quiz</td>
<td>2.5%</td>
<td>2.3%</td>
</tr>
</tbody>
</table>

#quads with LearningResourceType
20.665 in 2013
20.448 in 2014
Distribution of LRMI terms across PLDs

<table>
<thead>
<tr>
<th>PLD</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#quads</td>
<td>#docs</td>
</tr>
<tr>
<td>brainpop.com</td>
<td>513090</td>
<td>15062</td>
</tr>
<tr>
<td>claz.org</td>
<td>439102</td>
<td>24434</td>
</tr>
<tr>
<td>merlot.org</td>
<td>218466</td>
<td>10641</td>
</tr>
<tr>
<td>teacherspayteachers.com</td>
<td>55755</td>
<td>18322</td>
</tr>
<tr>
<td>curriki.org</td>
<td>11444</td>
<td>322</td>
</tr>
<tr>
<td>teachersnotebook.com</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>thegateway.org</td>
<td>668</td>
<td>66</td>
</tr>
<tr>
<td>pomagalo.com</td>
<td>118</td>
<td>38</td>
</tr>
<tr>
<td>senteacher.org</td>
<td>84</td>
<td>20</td>
</tr>
</tbody>
</table>

In *CreativeWork* subset
21 distinct PLDs in 2013
33 distinct PLDs in 2014
Distribution of LRMI terms across PLDs

but also....
Distribution of LRMI terms across PLDs

2013

2014

…and…
Observed frequent errors in LRMI statements

- Syntactic errors
  - Capitalization errors
  - Missing Slashes

- Semantic errors
  - Schema violation
  - Undefined properties
  - Object/Data properties

- Misused properties
Conclusions

• Significant growth in LRMI adoption

• Amount of documents drops significantly for certain LRMI providers

• Current investigation limited to the *CreativeWork* subset

• Ongoing work on entire *CreativeWork* subset

• Knowledge base population: using markup to generate KB of educational entities
Thank you!

Davide Taibi
National Research Council of Italy
Institute for Educational Technology
davide.taibi@itd.cnr.it