Extraction and Visualization of Metadata Analytics for Multimedia Learning Object Repositories: The case of TERENA TF-media network

Kostas Vogias¹, Ilias Hatzakis¹, Nikos Manouselis², Peter Szegedi³

¹ Greek Research and Technology Network
² Agro-Know Technologies
³ TERENA TF-media

Workshop on Learning Object Analytics for Collections, Repositories and Federations, 9 April, 2013
Metadata analysis is not something new

- Ochoa, Xavier, Klerkx, Joris, Vandeputte, Bram, and Duval, Erik.
    - Made the first fully quantitative study in a large number of Learning Repositories that belongs to a large organization like Globe

- Neven, Filip and Duval, Erik.
  - Reusable Learning Objects: a Survey of LOM-Based Repositories.

- Zschocke, Thomas and Beniest, Jan and Paisley, Courtney and Najjar, Jehad and Duval, Erik.
  - The LOM application profile for agricultural learning resources of the CGIAR
    - studied the use of LOM for the indexing of learning resources

- Manouselis, N, Salokhe, G, Keizer, J, and Rudgard, S.
  - Towards a Harmonization of Metadata Application Profiles for Agricultural Learning Repositories.
    - Made an analysis of the metadata schemas used by Repositories including Agricultural Learning Resources.
Why extraction of metadata analytics is important when we are developing a learning portal

Which metadata schema is used by our content providers?

How the providers are using different elements of the schema?

On which metadata schema our learning portal should rely?

Can we provide services based on metadata elements such as Subject, Type, Format, Keywords, Title and Descriptions?

Portal design decisions

Which languages can our portal support?
Study Objectives

- To perform a quantified study on the different metadata schemas used by TERENA TF media network

- To propose a metadata schema on which the TERENA OER portal will be based

- To verify if metadata analytics can constitute a tool that can facilitate the development of a learning portal that is based on metadata records aggregated by various content providers
What Terena OER Project is

- A European level metadata aggregation portal for Open Educational Resources (primarily audiovisual contents, recorded lectures) collected and maintained by institutional and national content repositories of the Research & Education Community.

- Main objectives of the project
  - Create a broker for national learning resource organizations.
  - Bridge the gap between the national repositories and the emerging global repositories (e.g., GLOBE) by establishing a European level metadata repository (i.e. aggregation point) for the national repositories acting in the R&E community. The European level repository will be a metadata repository only, the content remains in its original content repository.
Which Data Providers

- Successfully harvested.

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>Records Harvested</th>
<th>Metadata Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch Collection</td>
<td>619</td>
<td>oai_dc</td>
</tr>
<tr>
<td>DSpace at University Of Latvia</td>
<td>1009</td>
<td>oai_dc</td>
</tr>
<tr>
<td>OBAA Repository</td>
<td>56</td>
<td>oai_dc</td>
</tr>
<tr>
<td>RiuNet: Repositori Institucional de la Universitat Politècnica de València</td>
<td>21902</td>
<td>oai_dc</td>
</tr>
<tr>
<td>SCAM Repository</td>
<td>7351</td>
<td>oai_lom</td>
</tr>
<tr>
<td>Material Audiovisual ofrecido por el Campus do Mar</td>
<td>978</td>
<td>oai_dc</td>
</tr>
<tr>
<td>wikiwijs</td>
<td>26054</td>
<td>oai_dc</td>
</tr>
<tr>
<td>Małopolskie Towarzystwo Genealogiczne</td>
<td>181</td>
<td>oai_dc</td>
</tr>
</tbody>
</table>

Select the ones that can be used for the first version of TERENA OER portal
Which Data Providers

- Successfully harvested.

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>Records Harvested</th>
<th>Metadata Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switch Collection</td>
<td>619</td>
<td>oai_dc</td>
</tr>
<tr>
<td>DSpace at University Of Latvia</td>
<td>1009</td>
<td>oai_dc</td>
</tr>
<tr>
<td>OBAA Repository</td>
<td></td>
<td>oai_dc</td>
</tr>
<tr>
<td>RiuNet: Repositori Institucional de la Universitat Politècnica de València</td>
<td>21902</td>
<td>oai_dc</td>
</tr>
<tr>
<td>SCAM Repository</td>
<td></td>
<td>oai_lom</td>
</tr>
<tr>
<td>Material Audiovisual ofrecido por el Campus do Mar</td>
<td>978</td>
<td>oai_dc</td>
</tr>
<tr>
<td>wikiwijs</td>
<td>26054</td>
<td>oai_dc</td>
</tr>
<tr>
<td>Małopolskie Towarzystwo Genealogiczne</td>
<td>181</td>
<td>oai_dc</td>
</tr>
</tbody>
</table>

58.000 + instances

Select the ones that can be used for the first version of TERENA OER portal
HOW AND WHAT WE USED
HOW

• Repository Based Analysis.
• Metrics
  – **Element Completeness**: The percentage of records an element is present.
  – **Relative Entropy**: The amount of information an element carries.
  – **Vocabulary values distribution**: Format, Language and Type
  – **Language properties**: Attribute value usage frequency e.g. lang in free text elements Title, Description and Keyword

• The analysis was performed for a core set of metadata elements that is present in the studied repositories
• Use of a standard set of mappings from DC to LOM
What we have used

- ARIADNE Harvester

- A metadata analysis tool
  - Implemented using JAVA.
  - Metadata schema agnostic.
  - Metadata Analysis schemes:
    1. Repository based.
    2. Federation based.
  - Element based analysis:
    - Frequency
    - Completeness
    - Dimensionality
    - Relative Entropy
    - Specific element vocabulary extraction and usage frequency
  - Attribute based analysis.
    - Attribute value frequency
  - Input: XMLs
  - Output: CSV, TXT
Current version of the tool
Next version of the tool

Welcome to the metadata analytics module. Using this tool you can:
1. Compute several metrics for the metadata of your collection
2. Visualize the results
3. Save the results for further analysis

Check how it works
Start

Link to report a problem
Contact us
WHICH ELEMENTS ARE USED
<table>
<thead>
<tr>
<th>Element</th>
<th>Switch</th>
<th>UOL</th>
<th>OBAA</th>
<th>RiuNet</th>
<th>SCAM</th>
<th>Campus do Mar</th>
<th>wikiwis</th>
<th>Malopolskie</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>52.62</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>99.96</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Description</td>
<td>47.67</td>
<td>0.10</td>
<td>0.00</td>
<td>0.17</td>
<td>99.96</td>
<td>100.00</td>
<td>0.00</td>
<td>99.44</td>
</tr>
<tr>
<td>Subject</td>
<td>52.45</td>
<td>100.00</td>
<td>0.00</td>
<td>96.84</td>
<td>10.49</td>
<td>100.00</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Format</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>73.31</td>
<td>100.00</td>
<td>0.00</td>
<td>98.89</td>
</tr>
<tr>
<td>Identifier</td>
<td>52.62</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>97.24</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Language</td>
<td>3.10</td>
<td>95.34</td>
<td>0.00</td>
<td>99.91</td>
<td>92.45</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Type</td>
<td>52.33</td>
<td>99.31</td>
<td>0.00</td>
<td>96.57</td>
<td>74.12</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Publisher</td>
<td>47.38</td>
<td>44.84</td>
<td>0.00</td>
<td>43.97</td>
<td>65.43</td>
<td>100.00</td>
<td>100.00</td>
<td>97.78</td>
</tr>
<tr>
<td>Creator</td>
<td>52.45</td>
<td>92.76</td>
<td>89.09</td>
<td>99.67</td>
<td>65.43</td>
<td>100.00</td>
<td>0.00</td>
<td>32.78</td>
</tr>
<tr>
<td>Rights</td>
<td>52.62</td>
<td>0.20</td>
<td>0.00</td>
<td>100.00</td>
<td>89.27</td>
<td>100.00</td>
<td>96.07</td>
<td>0.00</td>
</tr>
<tr>
<td>Date</td>
<td>0.00</td>
<td>31.15</td>
<td>100.00</td>
<td>6.38</td>
<td>45.89</td>
<td>100.00</td>
<td>0.00</td>
<td>98.89</td>
</tr>
<tr>
<td>Relation</td>
<td>0.00</td>
<td>15.87</td>
<td>0.00</td>
<td>3.83</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>47.22</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.00</td>
<td>67.66</td>
<td>0.00</td>
<td>97.94</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
</tr>
<tr>
<td>Source</td>
<td>51.85</td>
<td>0.00</td>
<td>0.00</td>
<td>100.00</td>
<td>100.00</td>
<td>0.00</td>
<td>0.00</td>
<td>99.44</td>
</tr>
</tbody>
</table>
Define the mandatory elements for the schema of metadata aggregator.
WHICH VALUES ARE USED
Heterogeneity of information

Entropy distribution for vocabulary elements

<table>
<thead>
<tr>
<th>Format</th>
<th>Language</th>
<th>Type</th>
</tr>
</thead>
</table>

Switch | University Of Latvia | OBAA repository
RiuNet | SCAM | Campus do Mar
wikiwijs | Malopolskie
What type of digital objects

Of interest for TERENA OER

We need a filtering mechanism to keep only the LO that are suitable for TERENA
WHICH LANGUAGES ARE USED FOR THE ANNOTATION
Language distribution for description

- English: 8500
- Estonian: 3000
- Hungarian: 3000
- Spanish: 1000
- Greek: 500
- Deutsch: 500
- Norwegian: 500
- Russian: 500
- Romanian: 500
- Polish: 500
- Hindi: 100
- Czech: 100
- Slovenian: 100
- Latvian: 100
- Finnish: 100
Language distribution for keyword

English: 30000
Estonian: 3000
Hungarian: 1000
Spanish: 100
Greek: 10
Deutch: 1
Norwegian: 0
Russian: 0
Romanian: 0
Polish: 0
Hindi: 0
Czech: 0
Slovenian: 0
Latvian: 0
Finnish: 0
English can be the main language supported on the TERENA OER Portal
Conclusions

• Metadata Analysis helps in:
  – Defining the type of services that can be provided by the TERENA OER portal e.g. browse by type of LOs, elements that can be used for full text search
  – Providing recommendations back to the providers about usage of metadata elements
  – Validating the metadata records at harvesting time

• Next steps
  – Extend to more repositories of TERENA TF-media network
  – Combine with results of an online survey for content providers
  – Develop the web based version of the tool and provide it as an open source tool
Thank you!
Ilias Hatzakis, GRNET
hatzakis@admin.grnet.gr