ISA2: 2016.16 - Public Multilingual Knowledge Management Infrastructure for the Digital Single Market (PMKI)

Language Technology Industry Summit: 29/05/2018

Peter Schmitz, Enrico Francesconi, Najeh Hajlaoui, Brahim Batouche

Publications Office of the EU
Unit A2 – Common Data Repository
Section A2.003 – Repository
Agenda

- Presentation of the PMKI project
- Current status (May 2018)
- Collaboration and Communication
- Conclusions
- Annexes
# PMKI organisation

<table>
<thead>
<tr>
<th>Type of activity</th>
<th>Common services, common frameworks (Creation of a Public Multilingual Knowledge Infrastructure (PMKI))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service in charge</td>
<td>Publications Office of the European Union OP.A2</td>
</tr>
</tbody>
</table>
| Associated Services |  - European Commission  
  - DG CNECT.G3  
  - DIGIT.D2  
  - DGT.R3  
  - European Parliament  
  - DG Traduction, Terminology Coordination unit                                               |
| First approval of the proposal by the ISA² committee | March 2nd 2016 in the scope of the general presentation of the ISA² programme |
| Timeframe         | 2016 –2019                                                                                   |
Objectives of the PMKI project

• To create a proof-of-concept for a public multilingual knowledge infrastructure to enable interoperability between existing multilingual terminologies that of managed by different public entities and to make the data available for reuse by private and public entities.

• To propose a common data model for the representation of multilingual terminologies based on existing recognised standards.

• To develop semantic capabilities to enable the automated alignment of data from different sources in order to further increase their interoperability.
Context

- Digital Single Market (DSM) for Europe (priority of Junker's Commission)
  - Bringing down barriers, including language barriers
  - Unlock on-line, cross-border opportunities
- "Overcoming language barriers is vital for building the DSM, which is by definition multilingual"
  
  Blog post by Commissioner Andrus Ansip at the 27 May 2016

Contribution of PMKI:

Provision of a platform to made available interoperable public multilingual knowledge systems for reuse by different services that support the creation of a Multilingual Digital Single Market.
Positioning of PMKI in the common architecture for the multilingual DSM promoted by the EU Language Technology industry
Use-cases, overview

Localisation
... as support for systems that deal with the localisation of Websites

Machine translation:
... as support for machine translation systems (to increase quality)

Multilingual search
... as support for cross-lingual information retrieval solutions
(index and/or query extension)
Solution, PMKI Architecture

Users

ACCESS
SPARQL endpoint, RESTful API

KNOWLEDGE BASE
Concepts, relations

KNOWLEDGE MODEL
Core data model (Ontolex-Lemon)

SERVICES
Ingestion, alignment, data and knowledge model management, administration of the platform
Solution status

• Users, Access: dissemination platform (planned).

• Knowledge base: gold standard data, etc. (ongoing).

• Knowledge model: Standard representation and Core data model have been adopted (done).

• Services: VocBench – an open source collaborative platform for thesauri management – is adopted and adapted (ongoing).
VocBench

• VocBench is a collaborative and interactive system for the management of SKOS-XL thesauri.
• The new version of Vocbench (VB3) has been developed in the context of an ISA² action and is managed by the Publications Office, it has been used for editing the Eurovoc thesaurus.
• The core component for business and data access is offered by Semantic Turkey, an open source platform for knowledge management.
• It supports the search functionality based on string matching and queries on live triples using SPARQL with syntax completion and highlight.
## Milestones and deliverables for 2018

<table>
<thead>
<tr>
<th>Milestones</th>
<th>Status</th>
<th>Target date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1/2018: Technical architecture has been defined</td>
<td>Ongoing</td>
<td>Q4/18</td>
</tr>
<tr>
<td>Q2/2018: Start of feasibility study for multilingual indexing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q3/2018: Start of the development of the proof of concept</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q4/2018: Start of community creation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q4/2018: Study for the enhancement of the semantic capabilities of the platform has been finished</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deliverables</th>
<th>Status</th>
<th>Target date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic links</td>
<td>Ongoing</td>
<td>Q4/18</td>
</tr>
<tr>
<td>A core dataset with additional semantic links between the different language resources will be available.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beneficiaries</td>
<td>Communication channel</td>
<td>Activities</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>EU economy</strong></td>
<td>Web (information about the activity on the ISA² website, publicity on the Publications Office and other EU Institutions websites)</td>
<td>Information about the Project: Meetings with internal and external partners, Steering Committee meetings, Participation EGOVIS 2017 conference – Lyon – France Paper submission to EGOVIS 2018 conference in Regensburg, Germany</td>
</tr>
<tr>
<td><strong>EU language technology industry</strong></td>
<td>Web (information about the activity on the ISA² website, publicity on the Publications Office and other EU Institutions websites)</td>
<td>Contact with internal/external lge technology stakeholders</td>
</tr>
<tr>
<td></td>
<td>Conferences (delivery of presentations)</td>
<td>• Presentation of PMKI at the LT workshop 13/12/16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Acceptation of 2 papers (LREC 2018 &amp; AICOL2018)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• LT-Innovate Summit 2017,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Presentation of PMKI at ELRC conference 7-8/11/17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Presentation of PMKI at Meta-Forum 13-14/11/17</td>
</tr>
<tr>
<td><strong>Member States</strong></td>
<td>Web (information about the activity on the ISA² website, publicity on the Publications Office and other EU Institutions websites) Workshops (organisation of dedicated workshops with interested member states)</td>
<td>Collaboration and collection of use cases with:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Luxembourgish government (to align their vocabulary with EuroVoc)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• ITTIG-CNR, Florence, Italy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• BNL &quot;National Library of Luxembourg&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Participation to the 40 Years of MT conf., Grenoble, France</td>
</tr>
<tr>
<td><strong>EU Institutions</strong></td>
<td>Meetings Workshop (organisation of dedicated workshops with interested services)</td>
<td>• Meeting with EUI (European University Institute) – Florence-Italy to discuss possible collaboration on PMKI.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Meetings and contacts with DGT and DG-CONNECT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Participation to Language equality in the digital age, Towards a Human Language Project (10/01/2017) - EP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• HAEU &quot;Historical Archives of the European Union&quot; - Italy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Participation to the workshop on the Generation of Multilingual Parallel Documents (03/04/2017) – DGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Translating Europe Forum, Brussels 06-07/11/2017</td>
</tr>
<tr>
<td><strong>Terminology community</strong></td>
<td>Conferences (delivery of presentations)</td>
<td>• Contact and collaboration with EP-DG-Trad Term. Unit</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Accepted paper on terminology to AIDA Journal, Salarno - Italy</td>
</tr>
<tr>
<td><strong>Semantic Web community</strong></td>
<td>Conferences (delivery of presentations: SEMIC, dedicated conferences...)</td>
<td>• Participation to present an accepted paper at Ontolex2017 workshop to present the accepted paper, Galway, Ireland</td>
</tr>
</tbody>
</table>
VocBench-PMKI workshop

• Purpose: Presentation of the two strictly related ISA² actions (PMKI, VocBench)

• Participants:
  o Scientific community
  o Public Administrations
  o Language Technology Industries

• Agenda
  o Vocbench Workshop (1 day)
  o PMKI and Lemon-Ontolex Vocbench extension discussion (1 day)

• Planning:
  o Preparation: Q4/2017 – Q2/2018
  o Event: Q3/2018 (to be confirmed)
Conclusions

• Conceptual and technical work are well advanced
  o Data model has been validated by experts
  o Results of first automated alignments are promising

• Communication plan is respected
  o Participation in relevant conferences and workshops has been ensured and will continue
  o Dedicated PMKI Workshop in preparation

• First collaborations with external stakeholders have been established
  o Language technology industry
  o Collection of resources (contact with public organisations in Member States)
Questions?
Thank you for your attention
ISA² programme
You click, we link.
Stay in touch

ec.europa.eu/isa2

@EU_isa2  isa@ec.europa.eu

Run by the Interoperability Unit at DIGIT (European Commission) with 131€M budget, the ISA² programme provides public administrations, businesses and citizens with specifications and standards, software and services to reduce administrative burdens.
Annexes
Our Approach: Information Retrieval Framework

Source Thesaurus
(Ex: Eurovoc)

Target Thesaurus
(Ex: Eclais)
### Example 1 of alignment

<table>
<thead>
<tr>
<th>Properties</th>
<th>Concept 1 (EuroVoc)</th>
<th>Concept 2 (Eclas)</th>
<th>Similarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label (preflabel)</td>
<td>&quot;single market&quot;@en</td>
<td>&quot;EU/EC Single Market&quot;@en</td>
<td></td>
</tr>
<tr>
<td>Alternative Label</td>
<td>&quot;Community internal market&quot;@en</td>
<td>&quot;European internal market&quot;@en</td>
<td>0.47</td>
</tr>
<tr>
<td>(altlabel)</td>
<td>&quot;EC internal market&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;EU single market&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>definition</td>
<td>&quot;Refers to the EU as one territory without any internal borders or other regulatory obstacles to the free movement of goods and services as defined by the Treaty on the Functioning of the European Union.&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remark (scopeNote)</td>
<td>&quot;Do not confuse with EU market.&quot;@en</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Example 2 of alignment

<table>
<thead>
<tr>
<th>Properties</th>
<th>Concept 1 (EuroVoc)</th>
<th>Concept 2 (Eclas)</th>
<th>Similarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label (preflabel)</td>
<td>&quot;board of management&quot;@en</td>
<td>&quot;Executive management&quot;@en</td>
<td>0.31</td>
</tr>
<tr>
<td>Alternative Label (altlabel)</td>
<td>&quot;company management&quot;@en  &quot;management team&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>definition</td>
<td>&quot;Committee of managers responsible for managing the company and answerable to the board.&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remark (scopeNote)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 3 of alignment

<table>
<thead>
<tr>
<th>Properties</th>
<th>Concept 1 (EuroVoc)</th>
<th>Concept 2 (Eclas)</th>
<th>Similarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label (preflabel)</td>
<td>&quot;logistics&quot;@en</td>
<td>&quot;Equipment management&quot;@en</td>
<td></td>
</tr>
<tr>
<td>Alternative Label (altlabel)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>definition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remark (scopeNote)</td>
<td></td>
<td></td>
<td>0.0</td>
</tr>
</tbody>
</table>
Example 4 of alignment

<table>
<thead>
<tr>
<th>Properties</th>
<th>Concept 1 (EuroVoc)</th>
<th>Concept 2 (Eclas)</th>
<th>Similarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label (preflabel)</td>
<td>&quot;linguistics&quot;@en</td>
<td>&quot;Linguistics&quot;@en</td>
<td></td>
</tr>
<tr>
<td>Alternative Label (altlabel)</td>
<td>&quot;grammar&quot;@en &quot;lexicology&quot;@en &quot;phonetics&quot;@en &quot;pronunciation&quot;@en</td>
<td>&quot;Language study&quot;@en</td>
<td>0.23</td>
</tr>
<tr>
<td>definition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remark (scopeNote)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Example 5 of alignment

<table>
<thead>
<tr>
<th>Properties</th>
<th>Concept 1 (EuroVoc)</th>
<th>Concept 2 (Eclas)</th>
<th>Similarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label (preflabel)</td>
<td>&quot;market stabilisation&quot;@en</td>
<td>&quot;Market stabilisation&quot;@en</td>
<td></td>
</tr>
<tr>
<td>Alternative Label (altlabel)</td>
<td>&quot;improvement of market conditions&quot;@en</td>
<td>&quot;Market stabilisation&quot;@en</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;market regularisation&quot;@en</td>
<td>&quot;Language study&quot;@en</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;market regularization&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;market stabilization&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;stabilisation of prices&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;stabilization of prices&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>definition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remark (scopeNote)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Similarity measure: 0.49
Example 6 of alignment

<table>
<thead>
<tr>
<th>Properties</th>
<th>Concept 1 (EuroVoc)</th>
<th>Concept 2 (Eclas)</th>
<th>Similarity measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Label (preflabel)</td>
<td>&quot;press&quot;@en</td>
<td>&quot;Press&quot;@en</td>
<td></td>
</tr>
<tr>
<td>Alternative Label (altlabel)</td>
<td>&quot;journalism&quot;@en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>definition</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remark (scopeNote)</td>
<td></td>
<td></td>
<td>0.70</td>
</tr>
</tbody>
</table>
First results on internal alignment experiments

Data Quality

<table>
<thead>
<tr>
<th></th>
<th>Nbr concept</th>
<th>prefLabel</th>
<th>altLabel</th>
<th>definition</th>
<th>scopeNote</th>
<th>narrower</th>
<th>broader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eurovoc</td>
<td>6945</td>
<td>yes</td>
<td>yes</td>
<td>Yes partial</td>
<td>Yes partial</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>eclas</td>
<td>6392</td>
<td>yes</td>
<td>yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stw</td>
<td>6229</td>
<td>yes</td>
<td>yes</td>
<td>No (only 0.5%)</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Gold Standard

<table>
<thead>
<tr>
<th></th>
<th>Exact</th>
<th>Closer</th>
<th>Broder</th>
<th>Narrower</th>
<th>Related</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eurovoc-Eclas</td>
<td>3493</td>
<td>331</td>
<td>60</td>
<td>0</td>
<td>1369</td>
<td>5253</td>
</tr>
<tr>
<td>Eurovoc-Stw</td>
<td>2261</td>
<td>370</td>
<td>170</td>
<td>0</td>
<td>0</td>
<td>2803</td>
</tr>
</tbody>
</table>
### First results on internal alignment experiments

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Eclas (4099)</th>
<th>Stw(2959)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy %</strong></td>
<td>Eclas</td>
<td>Eurovoc</td>
</tr>
<tr>
<td>0.8</td>
<td>72.57</td>
<td>72.04</td>
</tr>
<tr>
<td><strong>Precision %</strong></td>
<td>99.80</td>
<td>99.79</td>
</tr>
<tr>
<td><strong>Recall %</strong></td>
<td>47.24</td>
<td>46.21</td>
</tr>
<tr>
<td><strong>F-measure %</strong></td>
<td>64.13</td>
<td>63.17</td>
</tr>
<tr>
<td>0.7</td>
<td>80.28</td>
<td>79.79</td>
</tr>
<tr>
<td>Precision %</td>
<td>98.96</td>
<td>99.01</td>
</tr>
<tr>
<td>Recall %</td>
<td>62.67</td>
<td>61.68</td>
</tr>
<tr>
<td>F-measure %</td>
<td>76.74</td>
<td>76.01</td>
</tr>
<tr>
<td>0.6</td>
<td>82.70</td>
<td>82.62</td>
</tr>
<tr>
<td>Precision %</td>
<td>98.89</td>
<td>98.96</td>
</tr>
<tr>
<td>Recall %</td>
<td>64.41</td>
<td>67.23</td>
</tr>
<tr>
<td>F-measure %</td>
<td>80.17</td>
<td>80.06</td>
</tr>
<tr>
<td>0.5</td>
<td>87.50</td>
<td>87.46</td>
</tr>
<tr>
<td>Precision %</td>
<td>98.32</td>
<td>98.38</td>
</tr>
<tr>
<td>Recall %</td>
<td>77.24</td>
<td>77.10</td>
</tr>
<tr>
<td>F-measure %</td>
<td>86.51</td>
<td>86.45</td>
</tr>
<tr>
<td>0.4</td>
<td>88.55</td>
<td>87.65</td>
</tr>
<tr>
<td>Precision %</td>
<td>90.43</td>
<td>90.30</td>
</tr>
<tr>
<td>Recall %</td>
<td>87.16</td>
<td>85.37</td>
</tr>
<tr>
<td>F-measure %</td>
<td>88.77</td>
<td>87.77</td>
</tr>
<tr>
<td><strong>Accuracy %</strong></td>
<td><strong>88.99</strong></td>
<td><strong>88.16</strong></td>
</tr>
<tr>
<td><strong>Precision %</strong></td>
<td><strong>88.36</strong></td>
<td><strong>88.15</strong></td>
</tr>
<tr>
<td><strong>Recall %</strong></td>
<td><strong>90.73</strong></td>
<td><strong>89.18</strong></td>
</tr>
<tr>
<td><strong>F-measure %</strong></td>
<td><strong>89.53</strong></td>
<td><strong>88.66</strong></td>
</tr>
<tr>
<td>0.3</td>
<td>88.63</td>
<td>87.72</td>
</tr>
<tr>
<td>Precision %</td>
<td>86.53</td>
<td>86.37</td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>