Towards a Public Multilingual Knowledge Management Infrastructure for the Digital Single Market (PMKI)

Peter Schmitz, Enrico Francesconi, Najeh Hajlaoui, Brahim Batouche

Publications Office of the EU - European Commission

OntoLex 2017 workshop - Galway - Ireland - 18/06/2017
Agenda

- Presentation of PMKI project
  - Context
  - Use cases

- PMKI Status (end of March 2017)
  - Deliverables
  - Milestones

- Collaboration and Communication

- Conclusion
# PMKI in short

<table>
<thead>
<tr>
<th>Type of Activity</th>
<th>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</td>
</tr>
</tbody>
</table>
| Associated Services | • DG Connect, DG DIGIT, DG DGT, Centre de Traduction  
                        • European Parliament: DG TRAD-Terminology Coordination unit |
| Approval of the proposal by the ISA² committee | March 2\textsuperscript{nd} 2016  
                                                   in the scope of the general presentation of the ISA² programme |
| Timeframe       | May 2016 - June 2019                                          |
Context

Digital Single Market for Europe (priority of Junker’s Commission)
- Bringing down barriers, including language barriers
- Unlock on-line, cross-border opportunities

Situation
- EU cross-border on-line services represent only 4% of the global Digital Market
- Only 7% of SMEs in the EU are actually selling cross-border

Actions: PMKI to support
- The implementation of interoperability between language resources through
  - Multilingual tools
  - Semantic Web technologies
in order to overcome language and semantic barriers in on-line services
The PMKI project

- PMKI is a ISA2 pilot project aiming to:
  - Create a proof-of-concept knowledge management infrastructure for language resources
  - Provide harmonization of their technical formats
  - Align concepts of different resources to facilitate interoperability and extensions
  - Set-up of a community and a governance structure allowing the integration of multilingual taxonomies/terminologies

- PMKI platform may represent a "one-stop-shop" language resources repository at European level.
Benefits of the PMKI action

- **Support for the development of multilingual digital tools**
  - Machine translation *(CEF Automated Translation Platform)*
  - Online service localisation
  - Multilingual search

- **PMKI vs ELRC** *(European Language Resource Coordination)*
  - ELRC aims **to identify and gather language and translation data**
  - PMKI aims **to harmonise multilingual language resources making them interoperable (creating links between them)**
PMKI use-case 1: content annotation

- **Semantic annotation** of digital contents
  - Word sense disambiguation
  - Multilingual alignment of terms in a context
  - Document classification
  - Semantic documents indexing

- **Benefits**
  - Multilingual Search
  - Machine Translation
  - Localisation
  - Multilingual comparative facilities
    (ex: Comparative Law)
PMKI use case-2: cross-lingual and cross-collection retrieval

- Accessing heterogeneous data sources in a distributed environment

- Language resources (thesauri or ontologies) can guarantee a better quality in document indexing (by controlled terms/concepts)

- Cross-collection and cross-lingual retrieval
  - providing queries from a single interface in a given language
  - retrieving pertinent documents from different collections and languages

- Quality of retrieval in single collections
  - linked to availability of specific thesauri

- Quality of retrieval in cross-collections
  - linked to interoperability among thesauri
PMKI use-case 3: Multilingual web sites and localisation

- Providing the **correct taxonomy in a given domain and in different languages**
- Extension of digital services
  - in a **new language**
  - in the **right context**

- Example:
  - multilingual localisation of a company website in 24 EU languages
PMKI use-case 4: Support to the LT industry

- **Language Technology** (Human Language Technology)
  - Natural Language Processing
  - Computational Linguistic
  - Speech Recognition

- PMKI will be a "one-stop-shop" language resources repository at the European level
PMKI use-case 5: Multilingual dictionary

- PMKI can be used as a pure translation dictionary
  - Providing a source and a service for looking up terms, translations, disambiguation, definitions, etc.
  - Allowing to browse available semantic networks (BabelNet, EuroVoc, etc.)

- Enable accessibility to information in other languages
PMKI use-case 6: support for MT & TM  "Transl. Memories"

- PMKI can be used as a data source contribution for Machine translation in all EU languages
  - Improving MT quality mainly for under-resourced language (Neural MT requires more data than Statistical MT)
    - Ex: adding EuroVoc to MT@EC systems
  - Providing filtered translation data to develop specific domain MT systems
    - Ex: specific needs for the translation in the legal domain as far as different languages and different legal systems is concerned
  - Enable MT domain adaptation

- PMKI can be connected to CAT "Computer Aided Translation" tools
  - To improve TM quality providing translation examples
  - To Help EU translators in their daily work providing dictionaries, thesauri, etc.
<table>
<thead>
<tr>
<th>Work Package</th>
<th>Description of milestones reached or to be reached</th>
<th>Milestone status</th>
<th>% Compl.</th>
<th>Start date</th>
<th>End date</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP0 - Project Management</td>
<td>Project organisation has been set-up</td>
<td>Done</td>
<td>100%</td>
<td>Q3/16</td>
<td>Q4/16</td>
</tr>
<tr>
<td>WP1 - Standards representation</td>
<td>Standard representation has been adopted</td>
<td>Done</td>
<td>100%</td>
<td>Q4/16</td>
<td>Q1/17</td>
</tr>
<tr>
<td>WP2 - PMKI core data model and extensions</td>
<td>Core data model and a first set of extensions have been defined (including documentation)</td>
<td>Ongoing</td>
<td>40%</td>
<td>Q4/16</td>
<td>Q2/17</td>
</tr>
<tr>
<td>WP3 - Design of the technical architecture for the PMKI platform</td>
<td>Technical architecture has been defined</td>
<td>Ongoing</td>
<td>50%</td>
<td>Q4/16</td>
<td>Q3/17</td>
</tr>
<tr>
<td>WP5 - Dissemination and government structure</td>
<td>Proposal for an adequate government structure has been defined</td>
<td>To do</td>
<td>0%</td>
<td>Q1/17</td>
<td>Q2/17</td>
</tr>
<tr>
<td>WP4 - Implementation and test of the technical infrastructure</td>
<td>First release of the system (operational proof of concept)</td>
<td>To do</td>
<td>0%</td>
<td>Q1/18</td>
<td>Q3/18</td>
</tr>
<tr>
<td>WP3 - Design of the technical architecture for the PMKI platform</td>
<td>Proposal for the implementation strategy</td>
<td>To do</td>
<td>0%</td>
<td>Q4/18</td>
<td>Q1/19</td>
</tr>
<tr>
<td>WP5 - Dissemination and government structure</td>
<td>Creation of the community</td>
<td>To do</td>
<td>0%</td>
<td>Q4/18</td>
<td>Q2/19</td>
</tr>
<tr>
<td>WP2 - PMKI core data model and extensions</td>
<td>Feasibility study for the enhancement of the semantic capabilities of the platform</td>
<td>Ongoing</td>
<td>5%</td>
<td>Q2/17</td>
<td>Q4/17</td>
</tr>
</tbody>
</table>
### PMKI Status: deliverables

<table>
<thead>
<tr>
<th>WP</th>
<th>Deliverables/Tasks</th>
<th>Status</th>
<th>S.date</th>
<th>E.date</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WP0: Project Management</strong></td>
<td>Project work plan (Excel, MS project)</td>
<td>done/ongoing</td>
<td>10/16</td>
<td>EoP</td>
</tr>
<tr>
<td></td>
<td>Project charter (PM2 template)</td>
<td>done</td>
<td>10/16</td>
<td>30/11/16</td>
</tr>
<tr>
<td></td>
<td>Progress tracking</td>
<td>done/ongoing</td>
<td>10/16</td>
<td>EoP</td>
</tr>
<tr>
<td></td>
<td>Day to day project management</td>
<td>done/ongoing</td>
<td>10/16</td>
<td>EoP</td>
</tr>
<tr>
<td></td>
<td>Business Case</td>
<td>done</td>
<td>01/2017</td>
<td>02/2017</td>
</tr>
<tr>
<td><strong>WP1: Standard representation</strong></td>
<td>Progress Report (10-11-12/2016)</td>
<td>done</td>
<td>01/2017</td>
<td>02/2017</td>
</tr>
<tr>
<td></td>
<td><strong>D1.1</strong> Detailed description of the work package - scope, content of the different deliverables (Report)</td>
<td>done</td>
<td>20/10/16</td>
<td>30/11/16</td>
</tr>
<tr>
<td></td>
<td><strong>D1.2</strong> Critical comparison of the available standards and recommendation (Report)</td>
<td>done</td>
<td>10/16</td>
<td>01/17</td>
</tr>
<tr>
<td><strong>WP2: PMKI core data model and extensions</strong></td>
<td><strong>D2.1</strong> Detailed description of the work package - scope, content of the different deliverables (Report)</td>
<td>done</td>
<td>20/10/16</td>
<td>30/11/16</td>
</tr>
<tr>
<td></td>
<td><strong>D2.2</strong> PMKI data model (Ontology based on RDF(S)/OWL technologies)</td>
<td>done</td>
<td>12/16</td>
<td>02/17</td>
</tr>
<tr>
<td></td>
<td><strong>D2.3</strong> Documentation of the PMKI data model (Report and online ontology documentation)</td>
<td>ongoing</td>
<td>02/17</td>
<td>04/17</td>
</tr>
<tr>
<td></td>
<td><strong>D2.4</strong> Analysis of the algorithms for language resources mapping (Report)</td>
<td>ongoing</td>
<td>02/17</td>
<td>05/17</td>
</tr>
<tr>
<td></td>
<td><strong>D2.5</strong> Feasibility study for the enhancement of the semantic capabilities of the platform</td>
<td>ongoing</td>
<td>Q2/17</td>
<td>Q4/17</td>
</tr>
<tr>
<td><strong>WP3: Design of the technical architecture for the PMKI platform</strong></td>
<td><strong>D3.1</strong> Detailed description of the work package - scope, content of deliverables (Report)</td>
<td>done</td>
<td>20/10/16</td>
<td>30/11/16</td>
</tr>
<tr>
<td></td>
<td><strong>D3.2</strong> Analysis of available platforms for managing language resources (Report)</td>
<td>done</td>
<td>11/16</td>
<td>01/17</td>
</tr>
<tr>
<td></td>
<td><strong>D3.3</strong> Analysis of the possible interaction with CEF AT platform (Report)</td>
<td>done (under validation)</td>
<td>02/17</td>
<td>03/17</td>
</tr>
<tr>
<td></td>
<td><strong>D3.4</strong> Technical architecture of the PMKI platform, including specifications for the implementation of the operational proof of concept (Report)</td>
<td>planned</td>
<td>04/17</td>
<td></td>
</tr>
</tbody>
</table>
WP1: Standard representation

- D1.1: detailed description of the work package - scope, content of the different deliverables (Report)

- D1.2 Critical comparison of the available standards and recommendation (Report)
  - Analysis of state of the art: critical comparison of the available standard representations adopted for describing multilingual resources.
  - Identification and recommendation of the most sophisticated advanced technology.
  - State of the art description
  - Criteria of comparison
  - Possible linguistic resources that PMKI will deal with: Controlled vocabulary, Glossary, Thesaurus, Lexicon, Taxonomy, Semantic Network
  - Conclusion: Semantic web representation (OWL/RDF technologies)
WP2: PMKI core data model and extensions

- **D2.1:** detailed description of the work package - scope, content of the different deliverables (Report)

- **D2.2 PMKI data model (Ontology based on RDF(S)/OWL technologies)**
  - Definition of a core data model based on the standard representation recommended on WP1 in order to:
    - facilitate the interoperability between different terminologies, i.e. through a shared set of metadata, and to
    - harmonise the representation of the data
  - Analysis of data model candidates (SKOS, LEMON, Ontolex, GOLD, etc.)
  - Representation of samples from the selected linguistic resources
    (Controlled vocabulary, Glossary, Thesaurus, Lexicon, Taxonomy, Semantic Network)

  - Conclusion: OntoLEX (SKOS, LEMON)
WP3: Design of the technical architecture for the PMKI platform

- D3.1: detailed description of the work package - scope, content of the different deliverables (Report)

- D3.2 Analysis of available platforms for managing language resources (Report)
  - Analysis of PMKI requirements
    - Edition of resources
    - User account requirements (System Administrator, Project Manager, Project users)
    - Import/export of resources (Multi-Format, Management of format, Multilingualism)
    - Alignment of resources
    - Usability and legal terms (License, free-to-use, open source)
  - Analysis and recommendation of available platforms for managing language resources
    - Criteria of the analysis: based on PMKI requirements
    - Examples of platforms (VocBench, BioPortal, BabelNet, etc.)
  - Conclusion: VocBench is preferred (mainly the next version)
D3.3 Analysis of the possible interactions with CEF.AT platform (Report)

- **Three levels of possible interactions**
  - **Strategic**
    - PMKI as a service for CEF.AT
    - PMKI as a language resource of MT (*More data for Neural MT than Statistical MT*)
    - Limitation of indirect translations
  - **Business**
    - Support for MT and TM "Translation Memories"
    - Support for thematic (MT and TM)
    - Support for EU translators
  - **Technical**
    - Support for Machine Aided/Assisted Translation (CAT) tools
    - Selection of data for better MT quality (*specific domain MT*)
    - Ontology based production of data for under-resourced languages

- **Conclusion:** PMKI can be very useful/helpful for CEF.AT
## Communication & Collaboration

<table>
<thead>
<tr>
<th>Beneficiaries</th>
<th>Communication channel</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>EU economy</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&lt;br&gt;• Meetings with internal and external partners&lt;br&gt;• Steering Committee meetings</td>
</tr>
<tr>
<td>EU language technology industry</td>
<td>Web (information about the activity on the ISA² website, publicity on the Publications Office and other EU Institutions websites) Conference (delivery of presentations)</td>
<td>Contact with internal/external language technology stakeholders&lt;br&gt;• Participation to brainstorming language technologies workshop (13/12/2016) DG-CONNECT&lt;br&gt;• LT-Innovate</td>
</tr>
<tr>
<td>Member States</td>
<td>Web (information about the activity on the ISA² website, publicity on the Publications Office and other EU Institutions websites)&lt;br&gt;Workshops (organisation of dedicated workshops with interested member states)</td>
<td>Collaboration and collection of use cases with:&lt;br&gt;• ITTIG-CNR, Florence, Italy&lt;br&gt;• BNL &quot;National Library of Luxembourg&quot;</td>
</tr>
<tr>
<td>EU Institutions</td>
<td>Meetings&lt;br&gt;Workshop (organisation of dedicated workshops with interested services)</td>
<td>• Meetings and contacts with DGT and DG-CONNECT&lt;br&gt;• Participation to Language equality in the digital age, Towards a Human Language Project (10/01/2017) - EP&lt;br&gt;• HAEU &quot;Historical Archives of the European Union&quot; - Italy&lt;br&gt;• Participation to the workshop on the Generation of Multilingual Parallel Documents (03/04/2017) - DGT</td>
</tr>
<tr>
<td>Terminology community</td>
<td>Conferences (delivery of presentations)</td>
<td>Contact and collaboration with EP-DG-Trad Terminology Unit</td>
</tr>
<tr>
<td>Semantic community</td>
<td>Web Conferences (delivery of presentations: SEMIC, dedicated conferences...)</td>
<td>Submission of paper to Ontolex2017 workshop</td>
</tr>
</tbody>
</table>
Conclusion

- PMKI contributes directly to implementing the European Interoperability Strategy (EIS)
  - It meets the recommendations included in the EIS
  - The creation of a PMKI will allow EU public administrations to create services that can be accessible and shareable independently from the language.
  - This action represents a good opportunity to harmonize the different language resources making them interoperable.

- Expected beneficiaries: EU economy, EU LT industry, Member States, EU Institutions, Terminology community, and Semantic Web community

- Synergies with external LT stakeholders will be considered
  - Verification and collecting use-cases