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
  - Milsetones
- Collaboration and Communication
- Conclusion

PMKI in short

Type of Activity	Creation of a **Public Multilingual Knowledge Infrastructure** (PMKI)
Service in charge	Publications Office of the European Union
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 2nd 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.

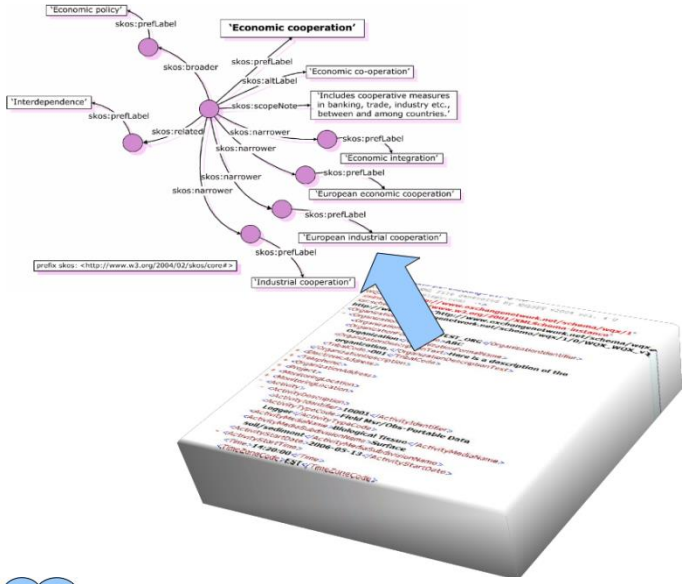
Benefits of the PMKI action

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

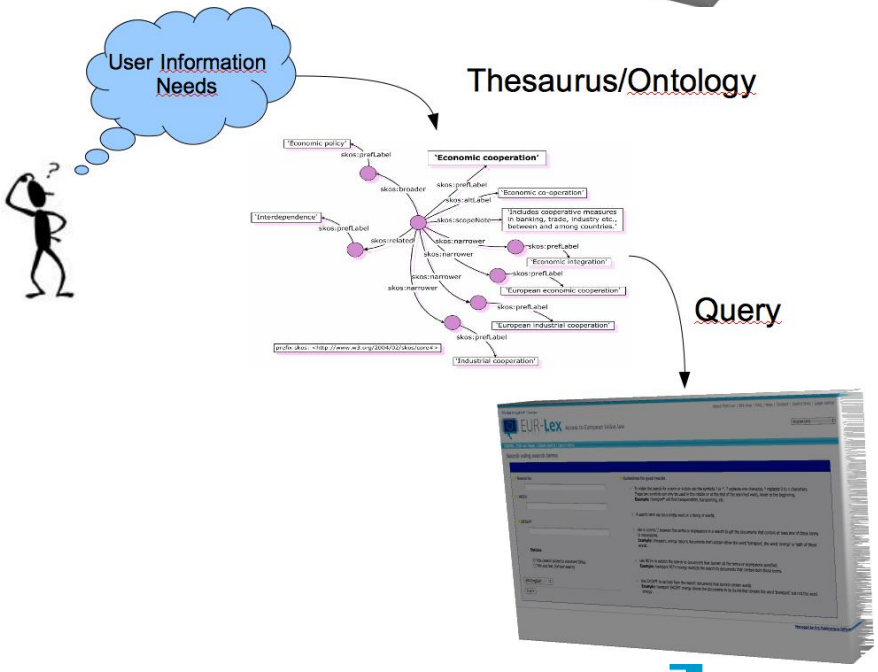
- **PMKI vs ELRC** (European Language Resource Coordination)
  - ELRC aims to identify and gather language and translation data
  - PMKI aims to harmonise multilingual language resources making them interoperable (creating links between them)

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

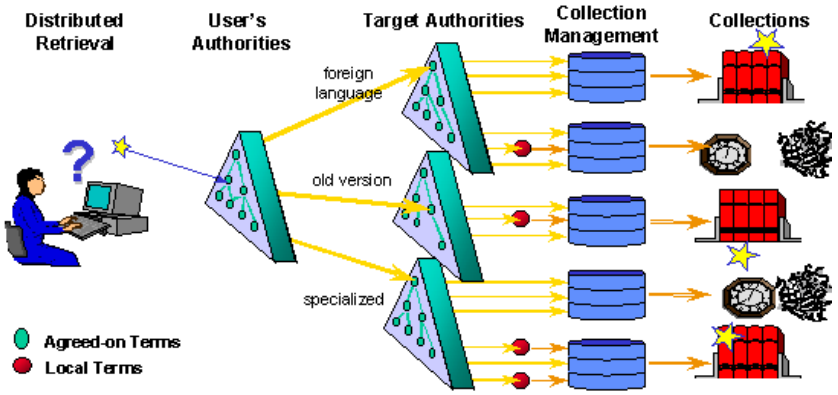


- Benefits
  - Multilingual Search
  - Machine Translation
  - Localisation
  - Multilingual comparative facilities (ex: Comparative Law)



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 compagny 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

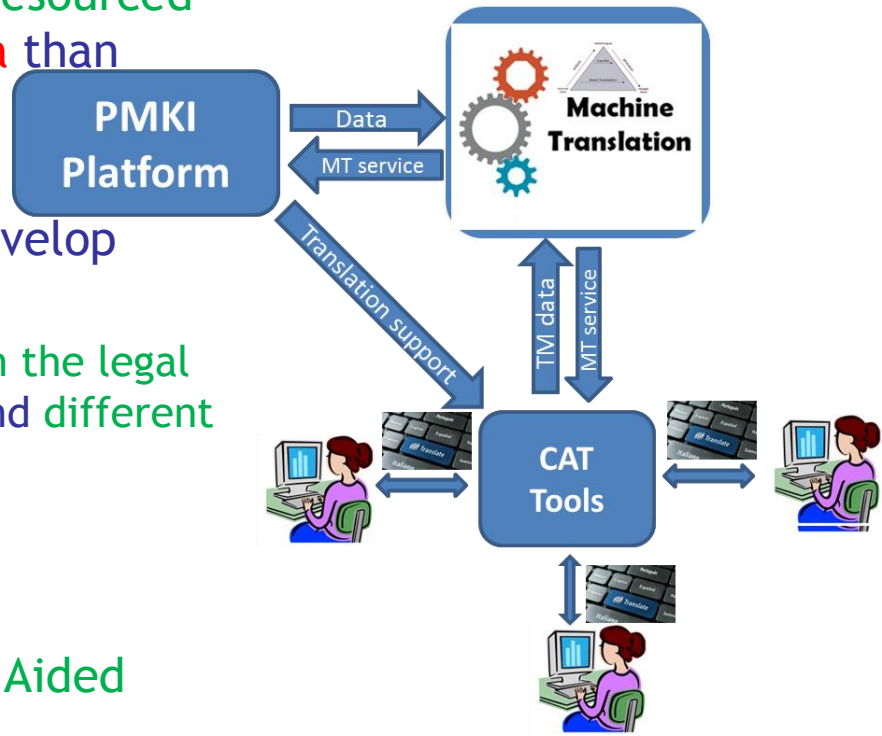
- PMKI will be a "one-stop-shop" language resources repository at the European level

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 avaliable 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.

PMKI Status: WPs and Milestones

Work Package	Description of milestones reached or to be reached	Milestone status	% Compl.	Start date	End date
WP0 – Project Management	Project organisation has been set-up	Done	100%	Q3/16	Q4/16
WP1 – Standards representation	Standard representation has been adopted	Done	100%	Q4/16	Q1/17
WP2 – PMKI core data model and extensions	Core data model and a first set of extensions have been defined (including documentation)	Ongoing	40%	Q4/16	Q2/17
WP3 - Design of the technical architecture for the PMKI platform	Technical architecture has been defined	Ongoing	50%	Q4/16	Q3/17
WP5 – Dissemination and government structure	Proposal for an adequate government structure has been defined	To do	0%	Q1/17	Q2/17
WP4 – Implementation and test of the technical infrastructure	First release of the system (operational proof of concept)	To do	0%	Q1/18	Q3/18
WP3 - Design of the technical architecture for the PMKI platform	Proposal for the implementation strategy	To do	0%	Q4/18	Q1/19
WP5 – Dissemination and government structure	Creation of the community	To do	0%	Q4/18	Q2/19
WP2 – PMKI core data model and extensions	Feasibility study for the enhancement of the semantic capabilities of the platform	Ongoing	5%	Q2/17	Q4/17

PMKI Status: deliverables

WP	Deliverables/Tasks	Status	S.date	E.date
WP0: Project Management	Project work plan (Excel, MS project)	done/ongoing	10/16	EoP
	Project charter (PM2 template)	done	10/16	30/11/16
	Progress tracking	done/ongoing	10/16	EoP
	Day to day project management	done/ongoing	10/16	EoP
	Business Case	done	01/2017	02/2017
	Progress Report (10-11-12/2016)	done	01/2017	02/2017
WP1: Standard representation	**D1.1** Detailed description of the work package - scope, content of the different deliverables (Report)	done	20/10/16	30/11/16
	D1.2 Critical comparison of the available standards and recommendation (Report)	done	10/16	01/17
WP2: PMKI core data model and extensions	**D2.1** Detailed description of the work package - scope, content of the different deliverables (Report)	done	20/10/16	30/11/16
	D2.2 PMKI data model (Ontology based on RDF(S)/OWL technologies)	done	12/16	02/17
	D2.3 Documentation of the PMKI data model (Report and online ontology documentation)	ongoing	02/17	04/17
	D2.4 Analysis of the algorithms for language resources mapping (Report)	ongoing	02/17	05/17
	D2.5 Feasibility study for the enhancement of the semantic capabilities of the platform	ongoing	Q2/17	Q4/17
WP3: Design of the technical architecture for the PMKI platform	**D3.1** Detailed description of the work package - scope, content of deliverables (Report)	done	20/10/16	30/11/16
	D3.2 Analysis of available platforms for managing language resources (Report)	done	11/16	01/17
	D3.3 Analysis of the possible interaction with CEF AT platform (Report)	Done (under validation)	02/17	03/17
	D3.4 Technical architecture of the PMKI platform, including specifications for the implementation of the operational proof of concept (Report)	planned	04/17	07/17

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)

WP3: Design of the tech. archi. for the PMKI plat.

- 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)
- **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

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

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