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

Type of activity	Common services, common frameworks (Creation of a Public Multilingual Knowledge Infrastructure (PMKI))
Service in charge	Publications Office of the European Union OP.A2
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"

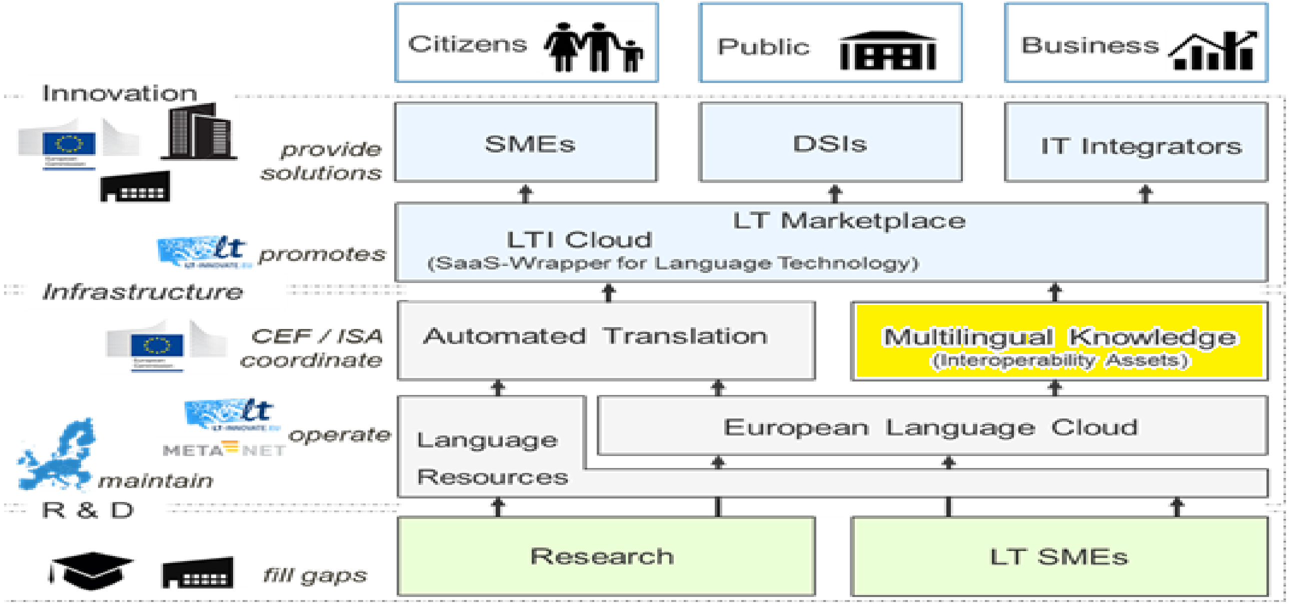
  Blog post by Commissioner Andrus Ansip at the 27 May 2016

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.

Context

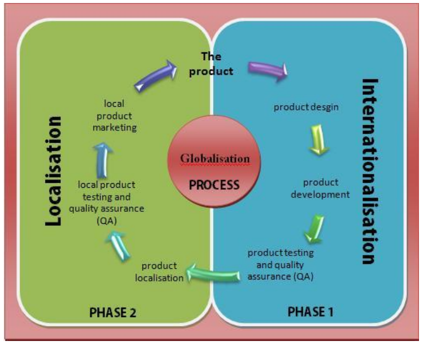
Positioning of PMKI in the common architecture for the multilingual DSM promoted by the EU Language Technology industry



Use-cases, overview

Localisation

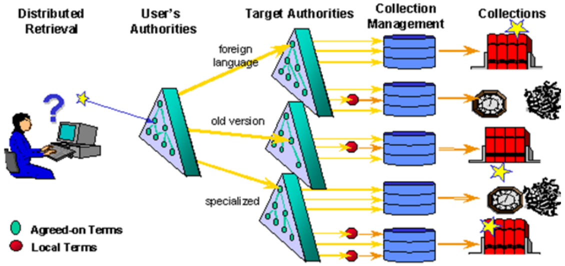
… as support for systems that deal with the localisation of Websites



Multilingual search

… as support for cross-lingual information retrieval solutions

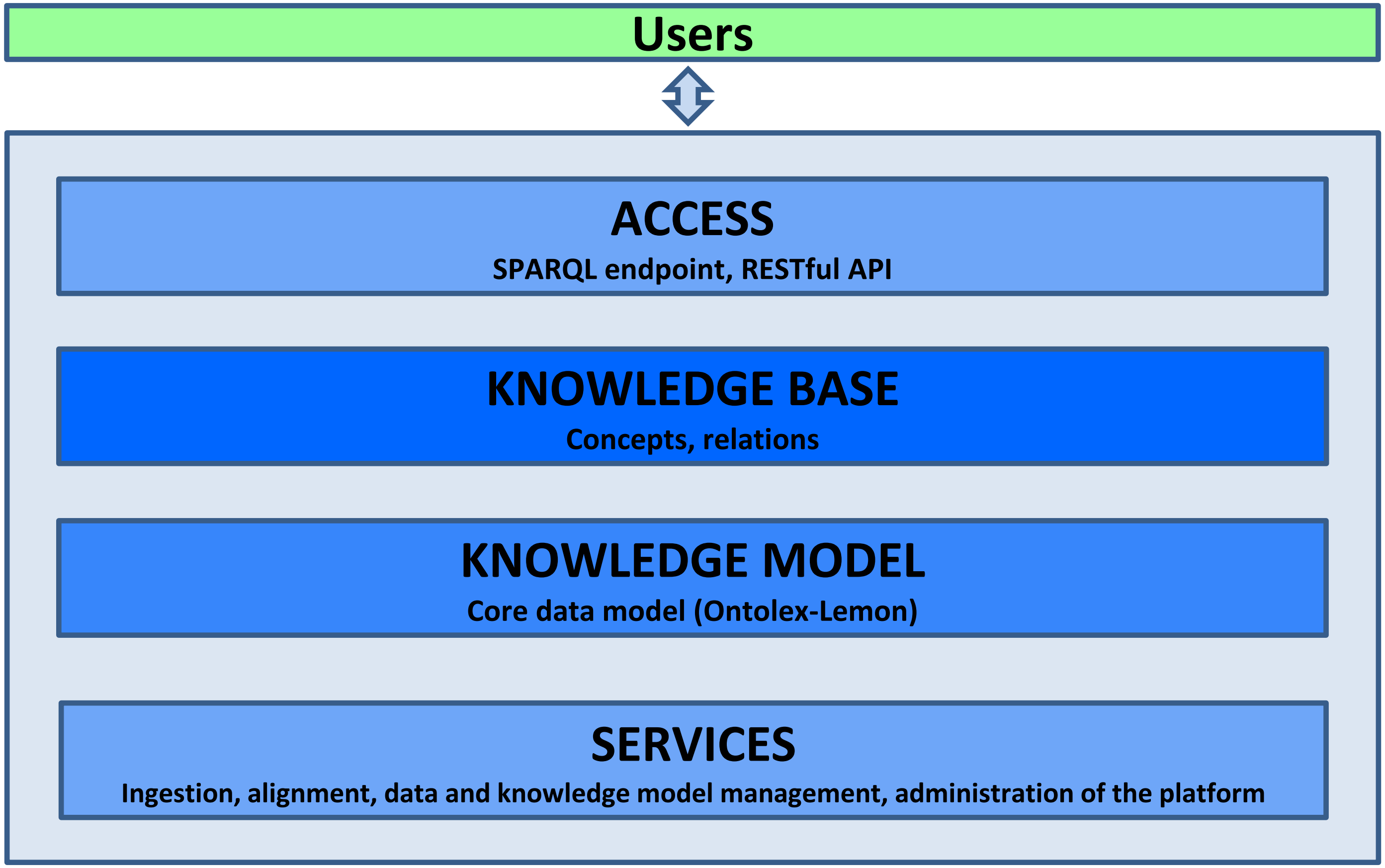
(index and/or query extension)



Machine translation:

…as support for machine translation systems (to increase quality)

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 ISA^2 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

Milestones
Q1/2018: Technical architecture has been defined
Q2/2018: Start of feasibility study for multilingual indexing
Q3/2018: Start of the development of the proof of concept
Q4/2018: Start of community creation
Q4/2018: Study for the enhancement of the semantic capabilities of the platform has been finished

Deliverables	Status	Target date
Semantic links A core dataset with additional semantic links between the different language resources will be available.	Ongoing	Q4/18

Communication & Collaboration (1)

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, Participation EGOVIS 2017 conference – Lyon – France Paper submission to EGOVIS 2018 conference in Regensburg, Germany
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 lge technology stakeholders • Presentation of PMKI at the LT workshop 13/12/16 • Acceptation of 2 papers (LREC 2018 & AICOL2018) • LT-Innovate Summit 2017, • Presentation of PMKI at ELRC conference 7-8/11/17 • Presentation of PMKI at Meta-Forum 13-14/11/17
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: • Luxemburgish government (to align their vocabulary with EuroVoc) • ITTIG-CNR, Florence, Italy • BNL "National Library of Luxembourg" • Participation to the 40 Years of MT conf., Grenoble, France
EU Institutions	Meetings Workshop (organisation of dedicated workshops with interested services)	• Meeting with EUI (European University Institute) – Florence-Italy to discuss possible collaboration on PMKI. • 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 • Translating Europe Forum, Brussels 06-07/11/2017
Terminology community	Conferences (delivery of presentations)	• Contact and collaboration with EP-DG-Trad Term. Unit • Accepted paper on terminology to AIDA Journal, Salarno - Italy
Semantic Web community	Conferences (delivery of presentations: SEMIC, dedicated conferences…)	• Participation to present an accepted paper at Ontolex2017 workshop to present the accepted paper, Galway, Ireland

Communication & Collaboration (2)

VocBench-PMKI workshop

- Purpose: Presentation of the two strictly related ISA^2 actions (PMKI, VocBench)
- Participants:
  - Scientific community
  - Public Administrations
  - Language Technology Industries
- Agenda
  - Vocbench Workshop (1 day)
  - PMKI and Lemon-Ontolex Vocbench extension discussion (1 day)
- Planning:
  - Preparation: Q4/2017 – Q2/2018
  - Event: Q3/2018 (to be confirmed)

Conclusions

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

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

- First collaborations with external stakeholders have been established
  - Language technology industry
  - Collection of resources (contact with public organisations in Member States)

Questions?

Thank you for your attention

European Commission

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.

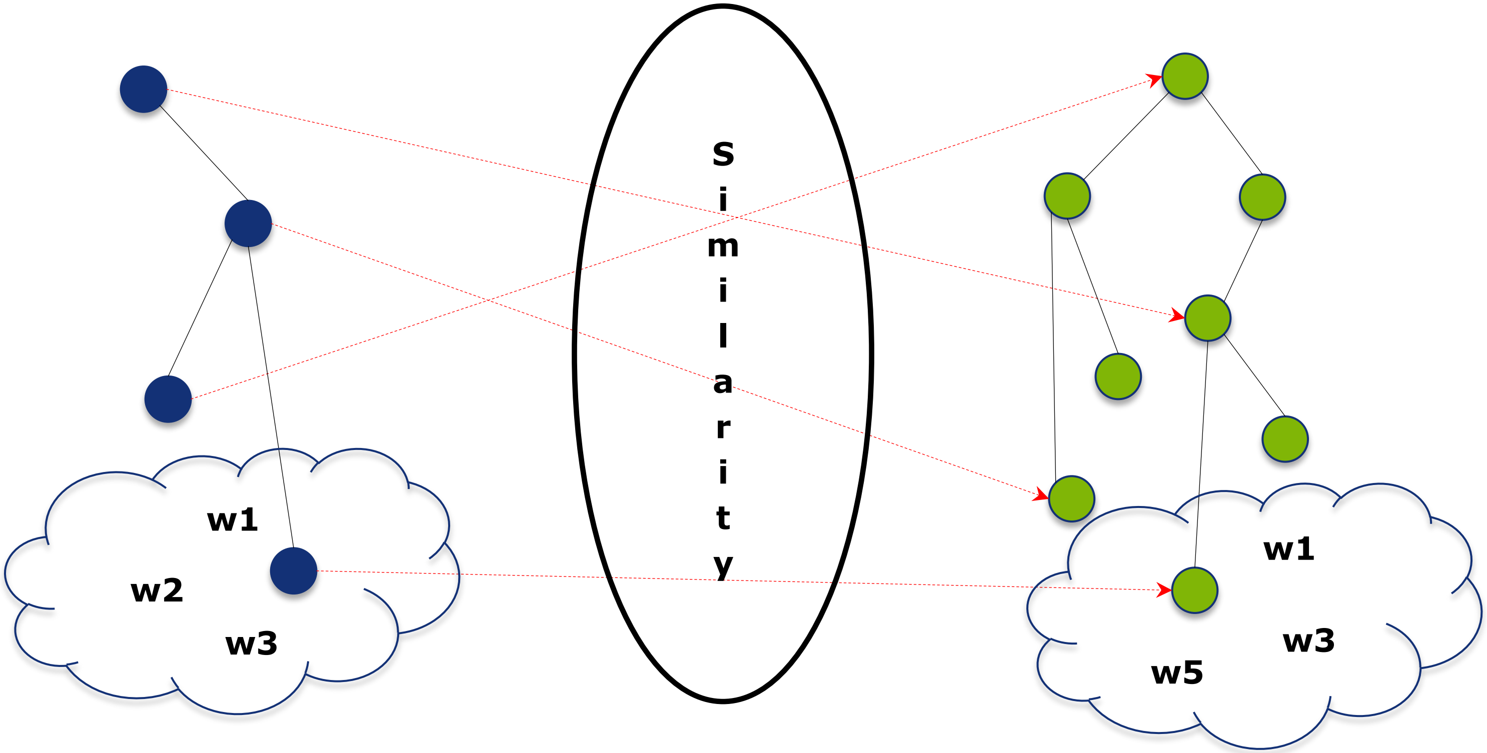
ISA²

Annexes

Our Approach: Information Retrieval Framework

Source Thesaurus
(Ex: Eurovoc)

Target Thesaurus
(Ex: Eclas)



Example 1 of alignment

Properties	Concept 1 (EuroVoc)	Concept 2 (Eclas)	Similarity measure
Identifier: uri	http://eurovoc.europa.eu/3299	http://ec.europa.eu/eclas/euc11/000000848	0.47
Label (preflabel)	**"single market"@en**	**"EU/EC Single Market"@en**	
Alternative Label (altlabel)	"Community internal market"@en "EC internal market"@en "EU single market"@en	"European internal market"@en	
definition	"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."@en		
Remark (scopeNote)	"Do not confuse with EU market."@en		

Example 2 of alignment

Properties	Concept 1 (EuroVoc)	Concept 2 (Eclas)	Similarity measure
Identifier: uri	http://eurovoc.europa.eu/447	http://ec.europa.eu/eclas/euc11/000001623	**0.31**
Label (preflabel)	**"board of management"@en**	**"Executive management"@en**	
Alternative Label (altlabel)	"company management"@en "management team"@en		
definition	"Committee of managers responsible for managing the company and answerable to the board."@en		
Remark (scopeNote)			

Example 3 of alignment

Properties	Concept 1 (EuroVoc)	Concept 2 (Eclas)	Similarity measure
Identifier: uri	http://eurovoc.europa.eu/1161	http://ec.europa.eu/eclas/euc11/000002579	**0.0**
Label (preflabel)	**"logistics"@en**	**"Equipment management"@en**	
Alternative Label (altlabel)			
definition			
Remark (scopeNote)			

Example 4 of alignment

Properties	Concept 1 (EuroVoc)	Concept 2 (Eclas)	Similarity measure
Identifier: uri	http://eurovoc.europa.eu/1670	http://ec.europa.eu/eclas/euc11/000003184	**0.23**
Label (preflabel)	"linguistics"@en	"Linguistics"@en	
Alternative Label (altlabel)	"grammar"@en "lexicology"@en "phonetics"@en "pronunciation"@en "semantics"@en "etymology"@en "spelling"@en	"Language study"@en	
definition			
Remark (scopeNote)			

Example 5 of alignment

Properties	**Concept 1 (EuroVoc)**	**Concept 2 (Eclas)**	**Similarity measure**
Identifier: uri	http://eurovoc.europa.eu/3165	http://ec.europa.eu/eclas/euc11/000005201	**0.49**
Label (preflabel)	"market stabilisation"@en	"Market stabilization"@en	
Alternative Label (altlabel)	"improvement of market conditions"@en "market regularisation"@en "market regularization"@en "market stabilization"@en "stabilisation of prices"@en "stabilization of prices"@en	"Language study"@en	
definition			
Remark (scopeNote)			

Example 6 of alignment

Properties	Concept 1 (EuroVoc)	Concept 2 (Eclas)	Similarity measure
Identifier: uri	http://eurovoc.europa.eu/2599	http://ec.europa.eu/eclas/euc11/000004299	**0.70**
Label (preflabel)	"press"@en	"Press"@en	
Alternative Label (altlabel)	"journalism"@en		
definition			
Remark (scopeNote)			

First results on internal alignment experiments

Data Quality

	Nbr concept	prefLabel	altLabel	definition	scopeNote	narrower	broader
Eurovoc	6945	yes	yes	Yes partial	Yes partial	yes	yes
eclas	6392	yes	yes	No	No	No	No
Stw	6229	yes	yes	No (only 0.5%)	No	No	No

Gold Standard

	Exact	Closer	Broder	Narrower	Related	Total
Eurovoc-Eclas	3493	331	60	0	1369	5253
Eurovoc-Stw	2261	370	170	0	0	2803

First results on internal alignment experiments

Target		Eclas (4099)		Stw(2959)	
Dictionary		Eclas	Eurovoc	stw	Eurovoc
0.8	Accuracy%	72.57	72.04	63.16	62.96
	Precision%	99.80	99.79	100.0	100.0
	Recall%	47.24	46.21	27.71	27.32
	F-measure%	64.13	63.17	43.40	42.91
0.7	Accuracy%	80.28	79.79	71.88	71.37
	Precision%	98.96	99.01	99.56	99.69
	Recall%	62.67	61.68	45.02	43.96
	F-measure%	76.74	76.01	62.00	61.02
0.6	Accuracy%	82.70	82.62	75.87	75.70
	Precision%	98.89	98.96	99.50	99.62
	Recall%	64.41	67.23	52.91	52.51
	F-measure%	80.17	80.06	69.09	68.77
0.5	Accuracy%	87.50	87.46	85.26	84.48
	Precision%	98.32	98.38	98.46	98.60
	Recall%	77.24	77.10	72.21	70.55
	F-measure%	86.51	86.45	83.32	82.25
0.4	Accuracy%	88.55	87.65	89.99	88.91
	Precision%	90.43	90.30	94.29	93.89
	Recall%	87.16	85.37	85.54	83.68
	F-measure%	88.77	87.77	89.70	88.49
0.3	**Accuracy%**	**88.99**	**88.16**	**91.68**	**90.87**
	Precision%	**88.36**	**88.15**	**91.45**	**91.65**
	Recall%	**90.73**	**89.18**	**92.30**	**90.31**
	F-measure%	**89.53**	**88.66**	**91.88**	**90.98**
0.2	Accuracy%	88.63	87.72	91.31	90.70
	Precision%	86.53	86.37	88.02	87.98